

STREAMLINING AND LARGE ANCESTRAL
GENOMES IN ARCHAEA INFERRRED WITH A
PHYLOGENETIC BIRTH-AND-DEATH MODEL

Supplemental Material

Miklós Csűrös Isván Miklós

July 4, 2009

Contents

I Supplemental Results	5
I.1 Evolutionary rates: correlations between sequence and gene content evolution	6
I.2 Evolutionary rates: uncertainty of estimates	6
I.3 Ancestral reconstruction: uncertainty of estimates	6
II Mathematical Framework for Phylogenetic Birth-And-Death Models	13
II.1 Introduction	14
II.2 Surviving lineages	16
II.3 Conditional likelihoods	17
II.4 Algorithm	19
II.5 Mathematical proofs	20
III Supplemental Methods	28
III.1 Universal conserved proteins	29
III.2 Likelihood correction for absent profiles	34
III.3 Inferring family sizes at ancestors and counting lineage-specific events	35
III.4 Rate variation	37

List of Figures

1	Rate comparisons for sequence and gene content evolution	7
2	Tree decomposition for computing posterior probabilities	36

List of Tables

1	Amino acid substitution and gene loss rates	8
2	Gene duplication and gain rates	9
3	Ancestral genome size estimates	10
4	Gene family losses and gains.	11
5	Gene family extensions and contractions.	12
6	Transient behavior of linear birth-and-death processes	15
7	Universal conserved proteins	29

Introduction

This Supplemental Material comprises three parts. Part I provides further details about some results in the main text. Part II describes the mathematical framework of phylogenetic birth-and-death models. Part III gives some additional information about the employed methods.

Part I

Supplemental Results

I.1 Evolutionary rates: correlations between sequence and gene content evolution

Figure 1 plots branch-specific rates $\hat{\mu}_e t_e$, $\hat{\lambda}_e t_e$, and $\hat{\kappa}_e t_e$ and expected numbers of substitutions for each branch e . In the figure, pairs of sibling terminal taxa are connected by lines.

"We found a conspicuous correlation across branches between the rate of sequence evolution and the component rates of gene content evolution."

I.2 Evolutionary rates: uncertainty of estimates

In order to assess the uncertainty of rate estimates and judge the significance of rate differences between sibling lineages, we used a bootstrapping procedure. Namely, we generated 100 random profile data sets by picking the same number of profiles, independently with replacement and uniform probability. On each bootstrap sample, we maximized the likelihood, and considered the minima and maxima of the inferred model parameters as the *bootstrap confidence intervals*. (The calculated bootstrap confidence interval thus corresponds to at least a 95% confidence interval over the true bootstrap distribution with probability $1 - 0.95^{100} = 0.994$). In an analogous manner, uncertainty in amino acid substitution rate estimates was computed using PhyML (Guindon and Gascuel, 2003) by generating 100 bootstrap samples from the multiple alignment of ribosomal proteins, and maximizing the likelihood on each one.

Table 1 shows amino acid substitution and loss rates, and table 2 shows duplication and gain rates.

"We examined the differences between evolutionary rates in sibling terminal taxa for signs of natural selection."

I.3 Ancestral reconstruction: uncertainty of estimates

As in Section I.2, 100 bootstrap samples were used to gauge the uncertainty in the inference of ancestral genome size and lineage-specific events. These latter were computed as posterior mean numbers for the profile dataset, with a correction for absent profiles. Table 3 shows the number of families present, and the number of multi-gene families at terminal and ancestral nodes. Tables 4 and 5 list the estimates for lineage-specific events of gene family loss, gain, expansion and contraction.

"We inferred a probable history of archaeal gene content using posterior probabilities for ancestral family sizes and family size changes"

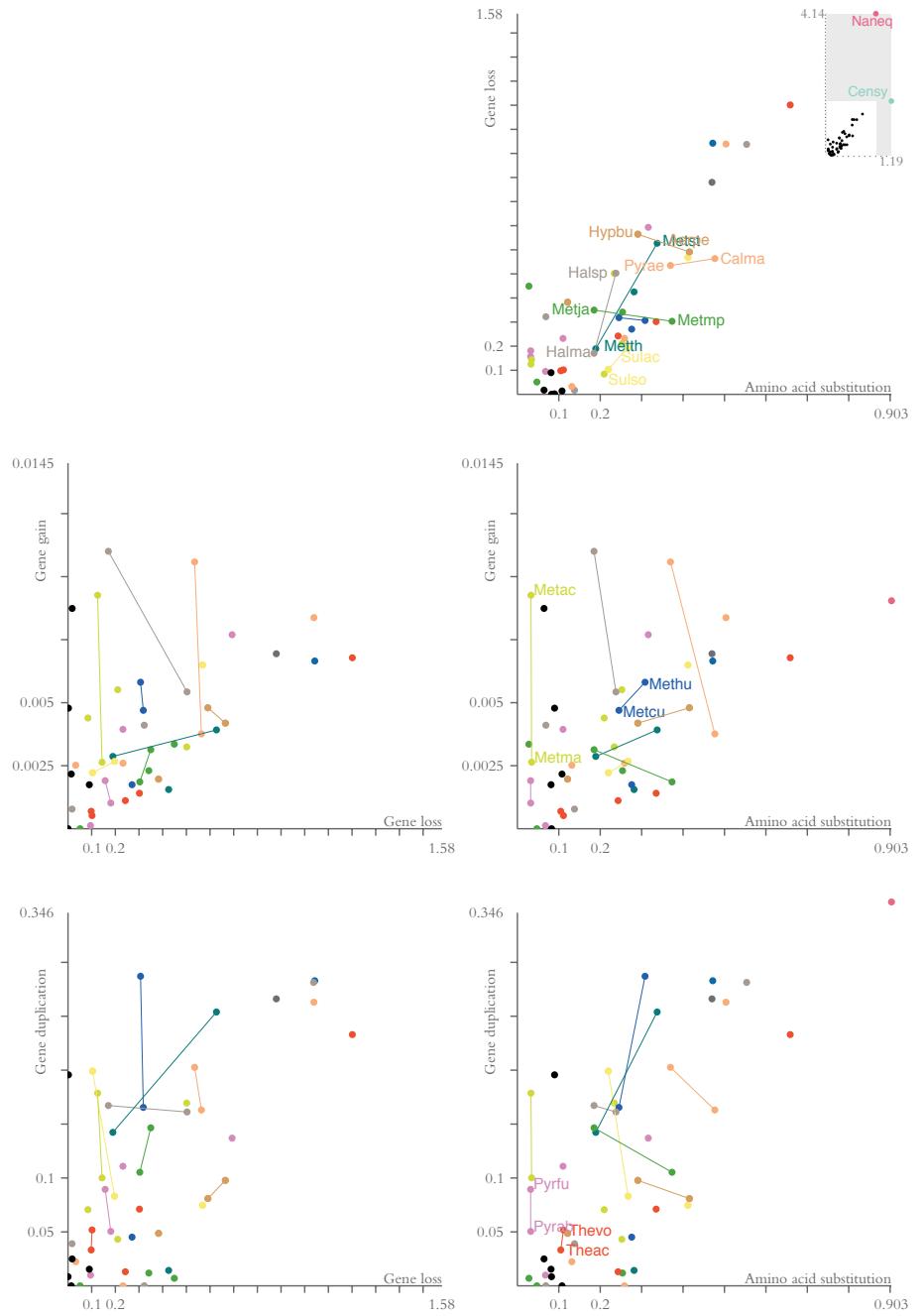


Figure 1: Rate comparisons for sequence and gene content evolution

Table 1: Amino acid substitution and gene loss rates. Substantially elevated rates (where confidence intervals do not overlap) between terminal sibling lineages are highlighted by boxes.

Branch	Substitution rate		Loss rate		Siblings level ^d
	estimate ^a	CI ^b	estimate ^c	CI ^b	
Aerpe	0.415	0.409–0.463	0.591	0.457–0.721	
Hypbu	0.291	0.269–0.326	0.665	0.531–0.819	
Desulfurococcales	0.121	0.11–0.144	0.383	0.253–0.479	
Sulso	0.22	0.206–0.256	0.104	0.0693–0.123	
Sulac	0.267	0.25–0.303	0.197	0.15–0.244	
Sulfolobales	0.412	0.385–0.462	0.569	0.455–0.678	
C1	0.123	0.106–0.15	0.00678	0.00175–0.0163	
Pyrae	0.37	0.352–0.427	0.535	0.384–0.704	
Calma	0.477	0.458–0.531	0.564	0.461–0.713	
Thermoproteaceae	0.259	0.237–0.297	0.233	0.15–0.345	
Thepe	0.504	0.478–0.558	1.04	0.849–1.3	
Thermoproteales	0.132	0.116–0.164	0.0326	0.00666–0.0696	
Naneq	0.903	0.884–0.999	4.14	3.34–4.8	
Pyrab	0.0321	0.0264–0.0431	0.181	0.142–0.231	
Pyrfu	0.0319	0.0255–0.0386	0.157	0.119–0.201	
Pyrococcus	0.0678	0.0553–0.0857	0.0952	0.0502–0.137	
Theko	0.11	0.0984–0.128	0.232	0.185–0.295	
Thermococcales	0.316	0.301–0.357	0.693	0.549–0.854	
Metka	0.472	0.445–0.523	1.04	0.853–1.24	
Metth	0.189	0.172–0.217	0.189	0.115–0.267	
Metst	0.337	0.322–0.379	0.627	0.491–0.792	
Methanobacteriales	0.282	0.266–0.318	0.425	0.329–0.542	
Metmp	0.373	0.356–0.421	0.303	0.224–0.372	
Metja	0.185	0.167–0.211	0.35	0.276–0.421	
Methanococcales	0.254	0.234–0.298	0.341	0.281–0.424	
M1	0.0471	0.0344–0.0678	0.0511	0.0142–0.0865	
Class I methanogens	0.0279	0.0159–0.0393	0.449	0.364–0.595	
Theac	0.105	0.0879–0.124	0.098	0.0657–0.127	
Thevo	0.111	0.0951–0.131	0.102	0.0608–0.127	
Thermoplasma	0.243	0.22–0.284	0.242	0.169–0.315	
Picto	0.335	0.314–0.379	0.302	0.221–0.39	
Thermoplasmatales	0.659	0.631–0.733	1.2	0.987–1.41	
Arcfu	0.47	0.448–0.523	0.88	0.729–1.07	
Metac	0.0328	0.0265–0.0411	0.125	0.102–0.156	
Metma	0.0343	0.0273–0.0472	0.144	0.1–0.181	
Methanosarcina	0.21	0.193–0.239	0.0842	0.0532–0.122	
Metbu	0.234	0.218–0.27	0.501	0.432–0.59	
Methanosarcinales	0.252	0.229–0.289	0.21	0.158–0.264	
Metcu	0.245	0.231–0.275	0.318	0.258–0.373	
Methu	0.308	0.29–0.343	0.307	0.228–0.355	
Methanomicrobiales	0.276	0.262–0.317	0.271	0.21–0.361	
Class II methanogens	0.0688	0.052–0.0912	0.322	0.244–0.422	
Halma	0.185	0.17–0.221	0.17	0.125–0.256	
Halsp	0.238	0.22–0.276	0.503	0.398–0.685	
Halobacteriales	0.554	0.533–0.61	1.04	0.836–1.22	
E5	0.138	0.123–0.168	0.0171	1.76e-05–0.0571	
E4	0.0812	0.0679–0.108	0.09	0.0163–0.147	
E3	0.0824	0.0636–0.102	0.000325	9.61e-06–0.0189	
E1	0.0643	0.0511–0.0844	0.0176	0.0109–0.0379	
E0	0.108	0.086–0.132	0.0139	7.64e-06–0.021	
Crenarchaeota	0.107	0.0854–0.143	0.108	0.063–0.153	
Euryarchaeota	0.0898	0.0618–0.122	0.002	0.000172–0.0307	
Censy	1.19	1.17–1.33	1.58	1.34–1.9	

^a maximum likelihood (ML) estimate on ribosomal proteins. ^b minimum and maximum of ML estimate in 100 bootstrap samples. ^c ML estimate of $\hat{\mu}_e \hat{t}_e$ on profile data set. ^d taxonomic group linking sibling terminal nodes.

Table 2: Gene duplication and gain rates. Substantially elevated rates (where confidence intervals do not overlap) between terminal sibling lineages are highlighted by boxes.

Branch	Duplication rate		Gain rate		Siblings level ^d
	estimate ^a	CI ^b	estimate ^c	CI ^b	
Aerpe	0.0809	0.0462–0.123	0.0048	0.0036–0.00573	
Hypbu	0.0978	0.0702–0.142	0.00419	0.00333–0.00532	
Desulfurococcales	0.0487	2.58e-06–0.102	0.00197	0.000531–0.00403	
Sulso	0.199	0.12–0.264	0.00222	0.00152–0.00299	
Sulac	0.0831	0.0652–0.108	0.00267	0.00211–0.00355	
Sulfolobales	0.0747	0.0367–0.115	0.0065	0.00498–0.00886	
C1	0.0273	3.43e-08–0.0904	0.00201	0.00112–0.00325	
Pyrae	0.203	0.128–0.28	0.0106	0.00924–0.0122	
Calma	0.163	0.116–0.213	0.00376	0.00317–0.00466	
Thermoproteaceae	2.36e-06	1.52e-06–0.0203	0.0026	0.00133–0.0048	family
Thepe	0.263	0.186–0.338	0.00837	0.00658–0.00991	
Thermoproteales	0.0223	8.76e-08–0.0837	0.00252	0.00148–0.00395	
Naneq	0.356	0.206–0.566	0.00904	0.00677–0.0115	
Pyrab	0.0505	0.0279–0.069	0.00102	0.000608–0.00139	
Pyrfu	0.0894	0.0558–0.114	0.00191	0.00143–0.0024	
Pyrococcus	0.00988	6.56e-07–0.0321	0.000125	7.19e-07–0.000607	genus
Theko	0.111	0.0872–0.133	0.00394	0.00312–0.00495	
Thermococcales	0.137	0.0871–0.178	0.0077	0.00555–0.00921	
Metka	0.283	0.216–0.334	0.00665	0.00543–0.00797	
Meth	0.142	0.0974–0.185	0.00287	0.00203–0.00344	
Mets	0.254	0.177–0.35	0.00392	0.00308–0.0049	
Methanobacteriales	0.0143	3.61e-06–0.0743	0.00155	0.000822–0.00273	
Metmp	0.105	0.0788–0.135	0.00186	0.00127–0.00245	
Metja	0.147	0.103–0.174	0.00313	0.00235–0.00409	
Methanococcales	0.0118	3.3e-06–0.0429	0.0023	0.00133–0.00343	
M1	5.17e-07	1.44e-07–0.0172	5.21e-07	1.45e-07–0.000108	
Class I methanogens	0.00698	3.73e-06–0.0363	0.00335	0.002–0.00595	
Theac	0.0332	0.0159–0.0448	0.000693	0.000436–0.00111	
Thevo	0.0518	0.0323–0.0735	0.00052	0.000308–0.00095	genus
Thermoplasma	0.013	1.99e-06–0.0439	0.00112	0.000533–0.00161	
Picto	0.0711	0.0457–0.105	0.00141	0.000803–0.00215	
Thermoplasmatales	0.233	0.117–0.318	0.00678	0.00448–0.00922	
Arcfu	0.266	0.2–0.333	0.00694	0.00453–0.00888	
Metac	0.179	0.148–0.21	0.00926	0.00758–0.0109	
Metma	0.1	0.0766–0.129	0.00264	0.00217–0.00339	genus
Methanosarcina	0.0706	0.0406–0.101	0.00439	0.00324–0.00582	
Metbu	0.17	0.145–0.225	0.00324	0.00247–0.00406	
Methanosarcinales	0.0432	0.00752–0.0748	0.00551	0.00433–0.00713	
Metcu	0.165	0.245–0.323	0.0047	0.00475–0.00669	
Methu	0.287	0.136–0.193	0.00581	0.0038–0.00591	
Methanomicrobiales	0.0452	0.0169–0.0674	0.00175	0.0011–0.00308	
Class II methanogens	7.15e-05	2.49e-06–0.0468	0.00411	0.00188–0.00649	
Halma	0.167	0.124–0.2	0.011	0.00866–0.0129	
Halsp	0.161	0.132–0.186	0.00543	0.00411–0.00658	
Halobacteriales	0.281	0.22–0.401	0.0203	0.0149–0.0265	
E5	0.0391	0.00172–0.0726	0.000779	5.21e-10–0.00225	
E4	0.0154	6.74e-07–0.0568	0.00175	8.15e-07–0.00366	
E3	0.00845	1.04e-10–0.0533	3.29e-09	9.72e-11–0.00058	
E1	0.025	1.5e-07–0.076	0.00874	0.00595–0.0119	
E0	1.41e-07	9.29e-11–0.155	0.00217	8.41e-11–0.00603	
Crenarchaeota	0.212	0.153–0.295	0.00455	0.0029–0.00611	
Euryarchaeota	0.196	0.0177–0.285	0.00478	0.0021–0.0077	
Censy	0.612	0.467–0.765	0.0196	0.0163–0.023	

^a maximum likelihood (ML) estimate on ribosomal proteins. ^b minimum and maximum of ML estimate in 100 bootstrap samples. ^c ML estimate of $\hat{\mu}_e \hat{t}_e$ on profile data set. ^d taxonomic group linking sibling terminal nodes.

Table 3: Gene count estimates and confidence intervals.

Node	Families present		Multi-gene families	
	estimate ^a	CI ^b	estimate ^c	CI ^b
Aerpe	1468	1393–1564	173	144–206
Hypbu	1380	1296–1475	180	146–211
Desulfurococcales	1840	1624–2030	284	202–351
Sulso	1908	1845–2025	416	375–467
Sulac	1789	1699–1887	273	244–320
Sulfolobales	1867	1757–1982	246	200–295
C1	2395	2188–2605	457	359–588
Pyrae	2052	1949–2163	304	264–351
Calma	1558	1487–1661	227	192–264
Thermoproteaceae	2101	1936–2477	277	201–382
Thepe	1487	1391–1568	231	196–267
Thermoproteales	2382	2131–2650	428	299–588
Naneq	516	474–598	29	20–43
Pyrab	1585	1511–1673	232	203–275
Pyrfu	1698	1627–1788	276	235–322
Pyrococcus	1743	1656–1845	235	185–279
Theko	1868	1778–1972	313	263–354
Thermococcales	1881	1724–2031	262	204–312
Metka	1423	1341–1504	188	157–214
Metth	1588	1525–1675	208	182–257
Metst	1286	1213–1362	171	143–199
Methanobacteriales	1566	1471–1716	104	70–149
Metmp	1495	1402–1584	169	138–205
Metja	1552	1464–1654	194	156–238
Methanococcales	1723	1565–1855	127	82–165
M1	2153	1905–2388	202	152–262
Class I methanogens	2271	1982–2458	224	162–293
Theac	1274	1193–1375	148	121–176
Thevo	1262	1190–1377	171	140–202
Thermoplasma	1311	1206–1399	139	111–172
Picto	1294	1225–1405	175	145–208
Thermoplasmatales	1472	1330–1637	200	142–246
Arcfu	1859	1771–1949	353	307–418
Metac	3111	2999–3229	737	681–800
Metma	2511	2396–2633	551	496–597
Methanosarcina	2635	2508–2748	462	399–547
Metbu	1812	1715–1889	318	278–357
Methanosarcinales	2469	2297–2670	331	260–409
Metcu	1928	1824–2041	389	337–425
Methu	2053	1964–2156	520	448–585
Methanomicrobiales	2063	1914–2251	287	226–352
Class II methanogens	2541	2346–2816	346	278–409
Halma	2911	2805–3020	683	617–743
Halsp	1980	1875–2083	374	328–439
Halobacteriales	2333	2193–2571	435	368–590
E5	3171	2836–3593	644	465–807
E4	3162	2867–3535	541	389–707
E3	3346	2926–3721	584	407–819
E1	3347	2926–3723	556	389–765
E0	2633	2356–3030	468	334–603
Crenarchaeota	2227	1996–2497	408	315–522
Euryarchaeota	2475	2116–2858	474	92–603
Censy	1545	1461–1657	213	187–260
LACA	2050	1909–2258	39	31–46

^a actual (for terminal nodes) or posterior mean number of families present at the node. ^b minimum and maximum of posterior mean or actual count in 100 bootstrap samples. ^c actual (for terminal nodes) or posterior mean number of multi-gene families.

Table 4: Gene family losses and gains.

Branch	Families lost		Families gained	
	estimate ^a	CI ^b	estimate ^c	CI ^b
Aerpe	650	471–793	278	213–329
Hypbu	692	504–846	232	185–280
Desulfurococcales	684	505–827	129	40–260
Sulso	145	95–175	186	132–256
Sulac	284	213–350	206	162–247
Sulfolobales	903	738–1085	375	297–468
C1	13	3–30	180	95–292
Pyrac	694	523–1034	645	575–730
Calma	766	625–1104	224	184–289
Thermoproteaceae	472	300–703	191	95–339
Thepe	1275	1096–1562	381	331–435
Thermoproteales	65	12–142	220	139–334
Naneq	2117	1775–2492	158	116–189
Pyrab	238	189–296	80	47–105
Pyrfu	197	150–250	152	121–182
Pyrococcus	148	76–208	10	0–46
Theko	308	241–383	295	236–372
Thermococcales	1166	947–1458	414	303–465
Metka	1154	930–1318	306	269–350
Metth	202	120–291	224	171–267
Metst	509	437–629	229	179–262
Methanobacteriales	686	490–854	99	54–177
Metmp	361	266–452	133	91–173
Metja	388	301–473	217	166–262
Methanococcales	586	457–734	156	85–207
M1	118	34–206	0	0–9
Class I methanogens	1280	995–1643	205	125–330
Theac	96	63–130	59	35–88
Thevo	93	59–125	44	26–76
Thermoplasma	243	162–337	83	40–112
Picto	278	215–358	100	57–140
Thermoplasmatales	2151	1739–2488	278	216–344
Arcfu	1643	1375–1967	339	248–414
Metac	272	205–328	748	652–821
Metma	333	235–404	209	171–260
Methanosarcina	198	123–294	364	283–452
Metbu	857	691–1013	200	153–255
Methanosarcinales	481	343–615	409	309–559
Metcu	465	373–562	330	291–390
Methu	426	319–515	416	352–505
Methanomicrobiales	603	449–802	124	78–206
Class II methanogens	904	656–1159	275	132–436
Halma	281	213–430	859	685–995
Halsp	688	578–900	335	256–398
Halobacteriales	1750	1419–2104	912	765–1139
E5	59	0–200	68	0–181
E4	324	58–526	141	0–300
E3	1	0–70	0	0–49
E1	46	28–99	760	552–995
E0	34	0–51	191	0–567
Crenarchaeota	199	116–291	376	242–496
Euryarchaeota	4	0–60	429	188–688
Censy	1255	1127–1444	749	673–832

^a posterior mean number of families lost on the branch leading to the node. ^b minimum and maximum of posterior mean in 100 bootstrap samples. ^c posterior mean number of families gained on the branch leading to the node.

Table 5: Gene family extensions and contractions.

Branch	Family expansions		Family contractions	
	estimate ^a	CI ^b	estimate ^c	CI ^b
Aerpe	82	48–108	34	16–63
Hypbu	82	56–114	50	30–68
Desulfurococcales	145	100–198	40	0–90
Sulso	21	11–34	174	135–226
Sulac	39	20–60	68	48–90
Sulfolobales	153	116–194	47	23–74
C1	5	1–10	50	3–140
Pyrae	85	57–117	109	76–145
Calma	80	49–115	80	46–120
Thermoproteaceae	125	83–182	3	1–25
Thepe	123	84–165	89	63–116
Thermoproteales	22	4–48	38	2–132
Naneq	42	11–57	13	6–22
Pyrab	37	23–52	38	23–51
Pyrfu	28	16–43	70	46–103
Pyrococcus	35	16–48	12	0–37
Theko	53	33–70	96	74–124
Thermococcales	156	107–186	89	44–122
Metka	70	50–94	85	64–108
Metth	15	7–26	109	75–144
Metst	36	23–56	84	57–118
Methanobacteriales	78	53–100	12	0–47
Metmp	28	16–42	71	46–93
Metja	33	19–48	95	67–113
Methanococcales	68	46–99	13	1–37
M1	21	6–37	0	0–32
Class I methanogens	224	160–289	11	2–46
Theac	13	7–20	23	9–35
Thevo	13	6–22	40	24–61
Thermoplasma	55	31–77	9	0–31
Picto	56	40–77	48	25–73
Thermoplasmatales	162	119–204	74	41–98
Arcfu	152	117–195	154	115–198
Metac	44	28–59	268	228–319
Metma	61	39–79	154	117–202
Methanosarcina	34	20–58	149	87–208
Metbu	93	65–121	127	103–161
Methanosarcinales	87	55–118	79	20–130
Metcu	59	36–78	154	126–194
Methu	46	31–62	231	201–283
Methanomicrobiales	103	80–135	65	27–95
Class II methanogens	219	146–288	7	3–72
Halma	46	34–80	236	172–291
Halsp	143	116–202	98	74–130
Halobacteriales	161	118–187	154	127–234
E5	15	0–46	116	7–197
E4	82	13–147	46	2–152
E3	0	0–21	29	0–165
E1	14	9–30	79	20–178
E0	12	0–20	6	0–333
Crenarchaeota	3	2–4	329	257–423
Euryarchaeota	0	0–1	375	46–477
Censy	11	7–15	114	91–144

^a posterior mean number of conserved families expanding (from size 1) on the branch leading to the node. ^b minimum and maximum of posterior mean in 100 bootstrap samples. ^c posterior mean number of conserved families contracting (to size 1) on the branch leading to the node.

Part II

Mathematical Framework for Phylogenetic Birth-And-Death Models

II.1 Introduction

This Part presents a standalone formulation of the mathematical framework, along with our results. For the sake of completeness, we reiterate the main definitions.

A *phylogenetic birth-and-death model* formalizes the evolution of an organism-specific census variable along a phylogeny. The phylogeny is a rooted tree, i.e., a connected acyclic graph in which the edges are directed away from a special node designated as the tree root; the terminal nodes, or *leaves*, are bijectively labeled by the organisms. The model specifies edge lengths, as well as birth-and-death processes (Ross, 1996; Kendall, 1949) acting on the edges. Let $\mathcal{E}(T)$ denote the set of edges, and let $\mathcal{V}(T)$ denote the node set of the tree. Populations of identical individuals evolve along the tree from the root towards the leaves by Galton-Watson processes. At non-leaf nodes of the tree, populations are instantaneously copied to evolve independently along the adjoining descendant edges. Let the random variable $\xi(x) \in \mathbb{N} = \{0, 1, 2, \dots\}$ denote the population count at every node $x \in \mathcal{V}(T)$. Every edge $xy \in \mathcal{E}(T)$ is characterized by a loss rate μ_{xy} , a duplication rate λ_{xy} and a gain rate κ_{xy} . If $(X(t): t \geq 0)$ is a linear birth-and-death process (Kendall, 1949; Takács, 1962) with these rate parameters, then

$$\mathbb{P}\left\{\xi(y) = m \mid \xi(x) = n\right\} = \mathbb{P}\left\{X(t_{xy}) = m \mid X(0) = n\right\},$$

where $t_{xy} > 0$ is the edge length, which defines the time interval during which the birth-and-death process runs. The joint distribution of $(\xi(x): x \in \mathcal{V}(T))$ is determined by the phylogeny, the edge lengths and rates, along with the distribution at the root ρ , denoted as $\gamma(n) = \mathbb{P}\{\xi(\rho) = n\}$. Specifically, for all set of node census values $(n_x: x \in \mathcal{V}(T))$,

$$\mathbb{P}\left\{\forall x \in \mathcal{V}(T): \xi(x) = n_x\right\} = \gamma(n_\rho) \prod_{xy \in \mathcal{E}(T)} w_{xy}[n_y | n_x] \quad (1)$$

where $w_{xy}[m|n] = \mathbb{P}\left\{X(t_{xy}) = m \mid X(0) = n\right\}$ denotes the transition probability on the edge xy for the Markov process operating there.

It is assumed that one can observe the population counts at the terminal nodes (i.e., leaves), but not at the inner nodes of the phylogeny. Since individuals are considered identical, we are also ignorant of the ancestral relationships between individuals within and across populations. The population counts at the leaves form a *phylogenetic profile*. Our central problem is to compute the likelihood of a profile, i.e., the probability of the observed counts for fixed model parameters.

The transient distribution of linear birth-and-death processes is well-characterized (Karlin and McGregor, 1958; Kendall, 1949; Takács, 1962), as shown in Table 6. Table 6 precisely states the distribution of xenolog and inparalog group sizes.

The distribution of population counts can be obtained analytically from the constituent distributions of Table 6, as shown by the following lemma.

Case	Condition	Transient distribution		Group
GLD	$\kappa > 0, \lambda > 0$	$\mathbb{P}\{X(t) = n\}$	$X(0) = 0\} = \text{NegativeBinomial}(n; \theta, q)$	xenolog
GL	$\kappa > 0, \lambda = 0$	$\mathbb{P}\{X(t) = n\}$	$X(0) = 0\} = \text{Poisson}(n; r)$	xenolog
DL	$\kappa = 0, \lambda > 0$	$\mathbb{P}\{X(t) = n\}$	$X(0) = 1\} = \text{ShiftedGeometric}(n; p, q)$	inparalog
PL	$\kappa = 0, \lambda = 0$	$\mathbb{P}\{X(t) = n\}$	$X(0) = 1\} = \text{Bernoulli}(n; 1 - p)$	inparalog

Parameters:

$$\begin{aligned} \theta &= \frac{\kappa}{\lambda} & r &= \kappa \frac{1 - e^{-\mu t}}{\mu} \\ p &= \frac{\mu - \mu e^{-(\mu-\lambda)t}}{\mu - \lambda e^{-(\mu-\lambda)t}} & \text{and} & \quad q = \frac{\lambda - \lambda e^{-(\mu-\lambda)t}}{\mu - \lambda e^{-(\mu-\lambda)t}} & \text{if } \lambda \neq \mu, \\ p &= q = \frac{\lambda t}{1 + \lambda t} & & & \text{if } \lambda = \mu. \end{aligned}$$

Distributions:

$$\begin{aligned} \text{NegativeBinomial}(n; \theta, q) &= \begin{cases} (1 - q)^\theta & \text{if } n = 0 \\ \frac{\theta(\theta+1)\cdots(\theta+n-1)}{n!} (1 - q)^\theta q^n & \text{if } n > 0 \end{cases} \\ \text{ShiftedGeometric}(n; p, q) &= \begin{cases} p & \text{if } n = 0; \\ (1 - p)(1 - q) & \text{if } n = 1 \\ (1 - p)(1 - q)q^{n-1} & \text{if } n > 1. \end{cases} \\ \text{Poisson}(n; r) &= e^{-r} \frac{r^n}{n!} \\ \text{Bernoulli}(n; 1 - p) &= \text{ShiftedGeometric}(n; p, 0) = \begin{cases} p & \text{if } n = 0 \\ 1 - p & \text{if } n = 1. \end{cases} \end{aligned}$$

Table 6: Transient behavior of linear birth-and-death processes with loss rate $\mu > 0$, gain rate κ and duplication rate λ : gain-loss-duplication (**GLD**), gain-loss (**GL**), duplication-loss (**DL**) and pure-loss (**PL**) models. The last column of the table shows the relevant group for computing transition probabilities in a phylogenetic birth-and-death model. For the meaning of xenolog and inparalog groups, see the main text.

Lemma 1. Let $(\zeta_i : i = 1, 2, \dots)$ be independent random variables that have identical, shifted geometric distributions with parameters p and q . Let η be a discrete nonnegative random variable that is independent from ζ_i , with probability mass function $\mathbb{P}\{\eta = m\} = H(m)$. Define $w[m|n] = \mathbb{P}\{\eta + \sum_{i=1}^n \zeta_i = m\}$ for all $m, n \geq 0$, and $w^*[m|n] = \mathbb{P}\{\eta + \sum_{i=1}^n \zeta_i = m; \forall \zeta_i > 0\}$ for all $m \geq n \geq 0$. These values can be expressed recursively as follows.

$$\begin{aligned} w[m|0] &= H(m) & \{m \geq 0\} & (2a) \\ w[0|n] &= p \cdot w[0|n-1] & \{n > 0\} & (2b) \\ w[1|n] &= p \cdot w[1|n-1] + (1-p)(1-q) \cdot w[0|n-1] & \{n > 0\} & (2c) \\ w[m|n] &= q \cdot w[m-1|n] \\ &\quad + (1-p-q) \cdot w[m-1|n-1] \\ &\quad + p \cdot w[m|n-1] & \{n > 0, m > 1\} & (2d) \end{aligned}$$

Furthermore,

$$\begin{aligned} w^*[m|0] &= H(m) & \{m \geq 0\} & (3a) \\ w^*[n|n] &= (1-p)(1-q) \cdot w^*[n-1|n-1] & \{n > 0\} & (3b) \\ w^*[m|n] &= q \cdot w^*[m-1|n] \\ &\quad + (1-p)(1-q) \cdot w^*[m-1|n-1] & \{m > n > 0\} & (3c) \end{aligned}$$

For every edge xy , Equation (2) provides the transition probabilities $w_{xy}[m|n] = w[m|n]$ in (1), when p , q and $H(m)$ are taken from Table 6 for the process operating on the edge xy . Equation (3) is used below in our formulas.

II.2 Surviving lineages

A key factor in inferring the likelihood formulas is the probability that a given individual at a tree node x has no descendants at the leaves within the subtree rooted at x . The corresponding *extinction probability* is denoted by D_x . An individual at node x is referred to as *surviving* if it has at least one progeny at the leaves descending from x . Let $\Xi(x)$ denote the number of surviving individuals at each node x . The distribution of $\Xi(x)$ can be related to that of $\xi(x)$ by

$$\mathbb{P}\{\Xi(x) = m\} = \sum_{i=0}^{\infty} \binom{m+i}{i} D_x^i (1-D_x)^m \mathbb{P}\{\xi(x) = m+i\}. \quad (4)$$

The next two lemmas characterize the number of surviving xenologs and inparalogs: they follow the same class of distributions as the total number of xenologs and inparalogs.

Lemma 2. For every edge $xy \in \mathcal{E}(T)$, let $G_y(n)$ denote the probability that there are n surviving members within an inparalog group at y . Then $G_y(n) =$

`ShiftedGeometric`($n; p', q'$) with

$$p' = \frac{p(1 - D_y) + (1 - q)D_y}{1 - qD_y} \quad \text{and} \quad q' = \frac{q(1 - D_y)}{1 - qD_y}.$$

Lemma 3. For every edge $xy \in \mathcal{E}(T)$, let $H_y(n)$ denote the probability that there are n xenologs at y that survive. If $\lambda_{xy} = 0$, then $H_y(n) = \text{Poisson}(n; r')$ where $r' = r(1 - D_y)$. If $\lambda_{xy} > 0$, then $H_y(n) = \text{NegativeBinomial}(n; \theta, q')$.

In the formulas to follow, we use the probabilities $w_y^*[m|n]$, which apply Lemma 1 to surviving populations on edge xy : $w_y^*[m|n] = w^*[m|n]$, where the latter is defined by Equation (3) with settings $p \leftarrow p'$, $q \leftarrow q'$, $H(m) \leftarrow H_{x_i}(m)$ from Lemmas 2 and 3.

Lemma 2 provides the means to compute extinction probabilities in a postorder traversal of the phylogeny.

Lemma 4. If x is a leaf, then $D_x = 0$. Otherwise, let x be the parent of x_1, x_2, \dots, x_c . Then D_x can be written as

$$D_x = \prod_{j=1}^c G_{x_j}(0). \quad (5)$$

II.3 Conditional likelihoods

Let $\mathcal{L}(T) \subset \mathcal{V}(T)$ denote the set of leaf nodes. A phylogenetic profile Φ is a function $\mathcal{L}(T) \mapsto \{0, 1, 2, \dots\}$, which are the population counts observed at the leaves. Define the notation $\Phi(\mathcal{L}') = (\Phi(x): x \in \mathcal{L}')$ for the partial profile within a subset $\mathcal{L}' \subseteq \mathcal{L}(T)$. Similarly, let $\xi(\mathcal{L}') = (\xi(x): x \in \mathcal{L}')$ denote the vector-valued random variable composed of individual population counts. The *likelihood* of Φ is the probability

$$L = \mathbb{P}\{\xi(\mathcal{L}(T)) = \Phi\}. \quad (6)$$

Let T_x denote the subtree of T rooted at node x . Define the *survival count range* M_x for every node $x \in \mathcal{V}(T)$ as $M_x = \sum_{y \in \mathcal{L}(T_x)} \Phi(y)$. The survival count ranges are calculated in a postorder traversal, since

$$M_x = \begin{cases} \Phi(x) & \text{if } x \text{ is a leaf} \\ \sum_{y \in \text{children}(x)} M_y & \text{otherwise.} \end{cases} \quad (7)$$

We compute the likelihood using *conditional survival likelihoods* defined as the probability of observing the partial profile within T_x given the number of surviving individuals $\Xi(x)$:

$$L_x[n] = \mathbb{P}\{\xi(\mathcal{L}(T_x)) = \Phi(\mathcal{L}(T_x)) \mid \Xi(x) = n\}.$$

For $m > M_x$, $L_x[m] = 0$. For values $m = 0, 1, \dots, M_x$, the conditional survival likelihoods can be computed recursively as shown in Theorem 5 below.

Theorem 5. If node x is a leaf, then

$$L_x[n] = \begin{cases} 0 & \text{if } n \neq \Phi(x); \\ 1 & \text{if } n = \Phi(x). \end{cases}$$

If x is an inner node with children x_1, \dots, x_c , then $L_x[n]$ can be expressed using $L_{x_i}[\cdot]$ and auxiliary values $A_{i;\cdot}$ and $B_{i;\cdot,\cdot}$ for $i = 1, \dots, c$ in the following manner. Let $w_{xx_i}^*[m|s]$ denote the transition probability in Lemma 1, applied to surviving individuals at x_i , using the distributions $H_{x_i}(\cdot)$ from Lemma 3 and $G_{x_i}(\cdot)$ from Lemma 2. Let $M[j] = \sum_{i=1}^j M_{x_i}$ for all $j = 1, \dots, c$ and $M[0] = 0$. Define also $D[j] = \prod_{i=1}^j G_{x_i}(0)$ and $D[0] = 1$. Auxiliary values $B_{i;t,s}$ are defined for all $i = 1, \dots, c$, $t = 0, \dots, M[i-1]$ and $s = 0, \dots, M_{x_i}$ as follows.

$$B_{i;0,s} = \sum_{m=0}^{M_{x_i}} w_{xx_i}^*[m|s] L_{x_i}[m] \quad \{0 \leq s \leq M_{x_i}\} \quad (8a)$$

$$B_{i;t,M_{x_i}} = G_{x_i}(0) B_{i;t-1,M_{x_i}} \quad \{0 < t \leq M[i-1]\} \quad (8b)$$

$$B_{i;t,s} = B_{i;t-1,s+1} + G_{x_i}(0) B_{i;t-1,s} \quad \left\{ \begin{array}{l} 0 \leq s < M_{x_i} \\ 0 < t \leq M[i-1] \end{array} \right\} \quad (8c)$$

For all $i = 1, \dots, c$ and $n = 0, \dots, M[i]$, define $A_{i;n}$ as

$$A_{1;n} = (1 - D[1])^{-n} B_{1;0,n}; \quad (9a)$$

$$A_{i;n} = (1 - D[i])^{-n} \sum_{\substack{0 \leq t \leq M[i-1] \\ 0 \leq s \leq M_{x_i} \\ t+s=n}} \text{Binomial}(s; n, D[i-1]) A_{i-1;t} B_{i;t,s}, \quad (9b)$$

where $i > 1$ in (9b).

For all $n = 0, \dots, M_x$, $L_x[n] = A_{c;n}$.

The complete likelihood is computed as

$$\begin{aligned} L &= \sum_{m=0}^{M_\rho} L_\rho[m] \mathbb{P}\{\Xi(\rho) = m\} \\ &= \sum_{m=0}^{M_\rho} L_\rho[m] \left(\sum_{i=0}^{\infty} \gamma(m+i) \binom{m+i}{i} D_\rho^i (1 - D_\rho)^m \right). \end{aligned} \quad (10)$$

For some parametric distributions γ , the infinite sum in (10) can be replaced by a closed formula for $\mathbb{P}\{\Xi(\rho) = m\}$. Theorem 6 below considers the stationary distributions for gain-loss-duplication and gain-loss models.

Theorem 6. For negative binomial or Poisson population distribution at the root, the likelihood can be expressed as shown below.

1. If $\gamma(n) = \text{Poisson}(n; r)$, then

$$\mathbb{P}\{\Xi(\rho) = m\} = \text{Poisson}(m; r') \quad (11a)$$

with $r' = r(1 - D_\rho)$.

2. If $\gamma(n) = \text{NegativeBinomial}(n; \theta, q)$, then

$$\mathbb{P}\{\Xi(\rho) = m\} = \text{NegativeBinomial}(m; \theta, q') \quad (11b)$$

$$\text{with } q' = \frac{q(1-D_\rho)}{1-qD_\rho}.$$

Consequently, the likelihood for a Poisson distribution at the root is computed as

$$L = \sum_{m=0}^{M_\rho} L_\rho[m] \text{Poisson}(m; \Gamma(1 - D_\rho)), \quad (12)$$

where Γ is the mean family size at the root.

II.4 Algorithm

The algorithm we describe computes the likelihood of a phylogenetic profile for a given set of model parameters. Algorithm COMPUTECONDITIONALS below proceeds by postorder (depth-first) traversals; the necessary variables are calculated from the leaves towards the root. The loop of Line 1 computes the transition probabilities $w^*[\cdot|\cdot]$, extinction probabilities D , and survival count ranges M . The loop of Line 9 carries out the computations suggested by Theorem 5.

Theorem 7. *Let T be a phylogeny with n nodes where every node has at most c^* children. Let h denote the tree height, i.e., the maximum number of edges from the root to a leaf. The COMPUTECONDITIONALS algorithm computes the conditional survival likelihoods for a phylogenetic profile Φ on T in $O(M^2h + c^*(Mh + n))$ time, where $M = M_\rho = \sum_x \Phi(x)$ is the total number of homologs.*

If c^* is constant, then the running time bound of Theorem 7 is $O(M^2h + n)$. For almost all phylogenies in a Yule-Harding random model, $h = O(\log n)$, so the typical running time is $O(M^2 \log n)$. For all phylogenies, $h \leq n - 1$, which yields a $O(M^2n)$ worst-case bound.

```

COMPUTECONDITIONALS
Input: phylogenetic profile  $\Phi$ 
1 for each node  $x \in \mathcal{V}(T)$  in a postorder traversal do
2   Compute the sum of gene counts  $M_x$  by (7).
3   Compute  $D_x$  using (5).
4   if  $x$  is not the root
5     then let  $y$  be the parent of  $x$ .
6     for  $n = 0, \dots, M_x$  do
7       for  $m = 0, \dots, M_x$  do
8         compute  $w_{yx}^*[m|n]$  using (3) with  $H_x(\cdot)$  and  $G_x(\cdot)$  from Lemmas 3 and 2.
9   for each node  $x \in \mathcal{V}(T)$  in a postorder traversal do
10    if  $x$  is a leaf
11      then for all  $n \leftarrow 0, \dots, \Phi(x)$  do set  $L_x[n] \leftarrow \{n = \Phi(x)\}$ 
12    else
13      Let  $x_1, \dots, x_c$  be the children of  $x$ 
14      Initialize  $M[0] \leftarrow 0$  and  $D[0] \leftarrow 1$ 
15      for  $i \leftarrow 1, \dots, c$  do
16        set  $M[i] \leftarrow M[i - 1] + M_{x_i}$  and  $D[i] \leftarrow D[i - 1] \cdot D_{x_i}$ 
17        for all  $t \leftarrow 0, \dots, M[i - 1]$  and  $s \leftarrow 0, \dots, M_{x_i}$  do
18          compute  $B_{i;t,s}$  by Eqs. (8)
19        if  $i = 1$  then for all  $n \leftarrow 0, \dots, M[i]$  do set  $A_{1;n} \leftarrow (1 - D[1])^{-n} B_{1;0,n}$ 
20        else
21          for all  $n \leftarrow 0, \dots, M[i]$  do initialize  $A_{i;n} \leftarrow 0$ 
22          for  $t \leftarrow 0, \dots, M[i - 1]$  and  $s \leftarrow 0, \dots, M_{x_i}$  do
23            set  $A_{i;n} \leftarrow A_{i;n} + \text{Binomial}(s; n, D[i - 1]) A_{i-1;t} B_{i;t,s}$ 
24          for all  $n \leftarrow 0, \dots, M[i]$  do  $A_{i;n} \leftarrow (1 - D[i])^{-n} A_{i;n}$ 
25        for all  $n \leftarrow 0, \dots, M_x$  do set  $L_x[n] \leftarrow A_{c;n}.$ 

```

II.5 Mathematical proofs

Proof of Lemma 1. Equations (2a) and (3a) are immediate since

$$w[m|0] = w^*[m|0] = \mathbb{P}\{\eta = m\} = H(m).$$

By the independence of ζ_i , for all $n > 0$,

$$w[0|n] = \mathbb{P}\{\eta + \sum_{i=1}^n \zeta_i = 0\} = w[0|n-1]\mathbb{P}\{\zeta_n = 0\} = w[0|n-1] \cdot p,$$

as in (2b).

Let $G(n) = \text{ShiftedGeometric}(n; p, q)$ be the common probability mass function of ζ_i . For $m, n > 0$,

$$\begin{aligned} w[m|n] &= \mathbb{P}\left\{\eta + \sum_{i=1}^n \zeta_i = m\right\} = \sum_{k=0}^m \mathbb{P}\{\zeta_n = k\} \cdot \mathbb{P}\left\{\eta + \sum_{i=1}^{n-1} \zeta_i = m - k\right\} \\ &= \sum_{k=0}^m G(k) \cdot w[m-k|n-1]. \quad (13) \end{aligned}$$

For $m = 1$, (13) is tantamount to (2c), since $G(0) = p$ and $G(1) = (1-p)(1-q)$. For $m > 1$, (13) can be further rewritten using $G(k) = qG(k-1)$ for all $k > 1$:

$$\begin{aligned} w[m|n] &= G(0) \cdot w[m|n-1] + G(1) \cdot w[m-1|n-1] \\ &\quad + \sum_{k=2}^m qG(k-1) \cdot w[m-k|n-1] \\ &= p \cdot w[m|n-1] + G(1) \cdot w[m-1|n-1] \\ &\quad + q \sum_{k=1}^{m-1} G(k) \cdot w[m-1-k|n-1] \\ &= p \cdot w[m|n-1] + G(1) \cdot w[m-1|n-1] \\ &\quad + q(w[m-1|n] - G(0) \cdot w[m-1|n-1]), \end{aligned}$$

which leads to the recursion of (2d) since $G(1) - qG(0) = 1 - p - q$.

Equation (3b) follows from

$$w^*[n|n] = \mathbb{P}\left\{\eta + \sum_{i=1}^n \zeta_i = n; \forall \zeta_i > 0\right\} = \mathbb{P}\{\eta = 0\} \cdot \prod_{i=1}^n \mathbb{P}\{\zeta_i = 1\} = H(0) (G(1))^n.$$

For $m > n > 0$,

$$\begin{aligned} w^*[m|n] &= \mathbb{P}\left\{\eta + \sum_{i=1}^n \zeta_i = m; \zeta_i > 0\right\} \\ &= \sum_{k=1}^{m-n+1} \mathbb{P}\{\zeta_n = k\} \cdot w^*[m-k|n-1] \\ &= \sum_{k=1}^{m-n+1} G(k) \cdot w^*[m-k|n-1] \\ &= G(1) \cdot w^*[m-1|n-1] + \sum_{k=2}^{m-n+1} qG(k-1) \cdot w^*[m-k|n-1] \\ &= G(1) \cdot w^*[m-1|n-1] + q \cdot w^*[m|n-1], \end{aligned}$$

as claimed in (3c). \square

Lemmas 2 and 3 rely on the following general result.

Lemma 8. *Let $\sigma \in \mathbb{R}$ be a fixed parameter, and let $\{a_n\}_{n=0}^\infty$ and $\{b_n\}_{n=0}^\infty$ be two number sequences related by the formula*

$$b_n = \sum_{i=0}^{\infty} \binom{n+i}{i} \sigma^i (1-\sigma)^n a_{n+i}. \quad (14a)$$

(We use the convention $0^0 = 1$ in the formula when $\sigma \in \{0, 1\}$.) Let $A(z) = \sum_n a_n z^n$ and $B(z) = \sum_n b_n z^n$ denote the generating functions for the sequences. Then

$$B(z) = A(\sigma + (1 - \sigma)z) \quad (14b)$$

Proof. If $\sigma = 0$, then $b_n = a_n$, and, thus (14b) holds. If $\sigma = 1$, then $b_0 = \sum_{k=0}^{\infty} a_k$, and $b_n = 0$ for $n > 0$, which implies (14b). Otherwise,

$$\begin{aligned} B(z) &= \sum_{n=0}^{\infty} z^n \sum_{m=n}^{\infty} a_m \binom{m}{n} \sigma^{m-n} (1-\sigma)^n \\ &= \sum_{m=0}^{\infty} a_m \sum_{n=0}^m z^n \binom{m}{n} \sigma^{m-n} (1-\sigma)^n \\ &= \sum_{m=0}^{\infty} a_m (\sigma + (1-\sigma)z)^m = A(\sigma + (1-\sigma)z), \end{aligned}$$

as claimed. \square

Corollary 9. Let $\{a_n\}_{n=0}^{\infty}$ and $\{b_n\}_{n=0}^{\infty}$ be two probability mass functions for non-negative integer random variables, related as in (14a).

1. If $a_n = \text{ShiftedGeometric}(n; p, q)$, then $b_n = \text{ShiftedGeometric}(n; p', q')$ with

$$p' = \frac{p(1-\sigma) + (1-q)\sigma}{1-q\sigma} \quad \text{and} \quad q' = \frac{q(1-\sigma)}{1-q\sigma}. \quad (15)$$

2. If $a_n = \text{NegativeBinomial}(n; \theta, q)$, then $b_n = \text{Negativebinomial}(n; \theta, q')$, where q' is defined as in (15).

3. If $a_n = \text{Poisson}(n; r)$, then $b_n = \text{Poisson}(n; r')$ with $r' = r(1-\sigma)$.

Proof. The corollary follows for plugging into Lemma 8 the generating functions $A(z) = \frac{p+z(1-p-q)}{1-qz}$, $A(z) = \left(\frac{1-q}{1-qz}\right)^{\theta}$ and $A(z) = e^{r(z-1)}$ for the shifted geometric, negative binomial, and Poisson distributions, respectively. \square

Proof of Lemma 2. Let ζ be the random variable that is the size of an inparalog group at node y . In order to have n surviving inparalogs, there must be $n+i$ inparalogs in total, out of which i do not survive, for some $i \geq 0$. Therefore,

$$G_y(n) = \sum_{i=0}^{\infty} \mathbb{P}\{\zeta = n+i\} \binom{n+i}{i} D_y^i (1-D_y)^n.$$

By Table 6, $\mathbb{P}\{\zeta = n\} = \text{ShiftedGeometric}(n; p, q)$ where the distribution parameters p and q are determined by the parameters of the birth-and-death process on the edge leading to y ($q = 0$ in the degenerate case where duplication rate is 0). The lemma thus follows from Corollary 9 with $\sigma = D_y$. \square

Proof of Lemma 3. Let η be the random variable that is the size of the xenolog group at y . In order to have n surviving xenologs, there must be $n+i$ xenologs in total, out of which i do not survive, for some $i \geq 0$. Therefore,

$$H_y(n) = \sum_{i=0}^{\infty} \mathbb{P}\{\eta = n+i\} \binom{n+i}{i} D_y^i (1-D_y)^n. \quad (16)$$

By Table 6 if $\lambda = 0$, then η has a Poisson distribution with parameter r ; otherwise, η has a negative binomial distribution with parameters θ and q . In either case, the lemma follows from Corollary 9 after setting $\sigma = D_y$. \square

Proof of Lemma 4. For leaves, the statement is trivial. When x is not a leaf, the lemma follows from the fact that survivals are independent between disjoint subtrees. \square

Proof of Theorem 5. The formulas are obtained by tracking survival within the lineages xx_i among the individuals at x . Define the indicator variables $X_{i,j}$ for each individual $j = 1, \dots, \xi(x)$ and lineage $i = 1, \dots, c$, taking the value 1 if and only if individual j has at least one surviving offspring at x_i . In order to work with the survivals, we introduce some auxiliary random variables that have the distributions of surviving xenologs and inparalogs. For every edge xx_i , define the sequence of independent random variables $(\zeta_{ij} : j = 1, 2, \dots)$ with identical distributions $G_{x_i}(\cdot)$, and the random variable η_i with distribution $H_{x_i}(\cdot)$. Consequently, $\mathbb{P}\{X_{i,j} = 0\} = \mathbb{P}\{\zeta_{ij} = 0\} = G_{x_i}(0)$, and $\mathbb{P}\{\Xi(x_i) = k \mid \xi(x) = n\} = \mathbb{P}\left\{\left(\eta_i + \sum_{j=1}^n \zeta_{ij}\right) = k\right\}$. By Lemma 2, $G_{x_i}(k) = \text{ShiftedGeometric}(k; p', q')$ with some p' and q' .

Define the shorthand notation $\Phi_i = \left\{\xi(T_{x_i}) = \Phi(T_{x_i})\right\}$ for observing the counts in the subtree rooted at x_i . Let

$$B_{i;t,s} = \mathbb{P}\left\{\Phi_i; \sum_{j=1}^s X_{i,j} = s \mid \xi(x) = t+s\right\}.$$

In other words, if there are $t+s$ individuals at x , then $B_{i;t,s}$ is the probability that s selected individuals survive in the lineage xx_i and the given profile is

observed in T_{x_i} . By Lemma 1,

$$\begin{aligned}
B_{i;0,s} &= \mathbb{P}\left\{\Phi_i; \sum_{j=1}^s X_{i,j} = s \mid \xi(x) = s\right\} \\
&= \sum_{m=0}^{M_{x_i}} \mathbb{P}\left\{\Phi_i; \Xi(x_i) = m; \sum_{j=1}^s X_{i,j} = s \mid \xi(x) = s\right\} \\
&= \sum_m \mathbb{P}\left\{\Phi_i \mid \Xi(x_i) = m; \xi(x) = s\right\} \cdot \mathbb{P}\left\{\Xi(x_i) = m; \sum_{j=1}^s X_{i,j} = s \mid \xi(x) = s\right\} \\
&= \sum_m \mathbb{P}\left\{\Phi_i \mid \Xi(x_i) = m\right\} \cdot \mathbb{P}\left\{\left(\eta_i + \sum_{j=1}^s \zeta_{ij}\right) = m; \forall j \leq s: \zeta_{ij} \neq 0\right\} \\
&= \sum_m L_{x_i}[m] \cdot w_{x_i}^*[m|s]
\end{aligned}$$

as shown in (8a). If $t > 0$, then

$$\begin{aligned}
B_{i;t,s} &= \mathbb{P}\left\{\Phi_i; X_{i,s+1} = 0; \sum_{j=1}^s X_{ij} = s \mid \xi(x) = t+s\right\} \\
&\quad + \mathbb{P}\left\{\Phi_i; X_{i,s+1} = 1; \sum_{j=1}^s X_{ij} = s \mid \xi(x) = t+s\right\} \\
&= G_{x_i}(0) \cdot \mathbb{P}\left\{\Phi_i; \sum_{j=1}^s X_{ij} = s \mid \xi(x) = t+s-1\right\} \\
&\quad + \mathbb{P}\left\{\Phi_i; \sum_{j=1}^{s+1} X_{ij} = s+1 \mid \xi(x) = t+s\right\} \\
&= G(0) \cdot B_{i;t-1,s} + B_{i;t-1,s+1},
\end{aligned}$$

which is tantamount to (8c). Equation (8b) follows from the fact that $B_{i;t,s} = 0$ for $s > M_{x_i}$.

For every i , let \mathcal{Y}_i denote the set of individuals at x that survive in at least one of the lineages x_1, \dots, x_i , i.e., $\mathcal{Y}_i = \left\{j: \{X_{1j} + \dots + X_{ij} \neq 0\}\right\}$. Given the exchangeability of individuals, $\mathbb{P}\left\{\Phi_1; \dots; \Phi_i \mid \mathcal{Y}_i = \mathcal{Y}\right\}$ is the same for all sets of the same size $|\mathcal{Y}| = n$. Let

$$A_{i;n} = \mathbb{P}\left\{\Phi_1; \dots; \Phi_i \mid |\mathcal{Y}_i| = n\right\}.$$

In particular, $|\mathcal{Y}_c| = \Xi(x)$ and, thus, $A_{c;n} = L_x[n]$.

Since $A_{i;n}$ is the same for all values of $\xi(x) \geq n$, and Φ_1 is determined solely

by survival in lineage xx_1 ,

$$\begin{aligned}
B_{1;0,n} &= \mathbb{P}\left\{\Phi_1; \sum_{j=1}^n X_{i,j} = n \mid \xi(x) = n\right\} \\
&= \mathbb{P}\left\{\sum_{j=1}^n X_{i,j} = n \mid \xi(x) = n\right\} \cdot \mathbb{P}\left\{\Phi_1 \mid \sum_{j=1}^n X_{i,j} = n; \xi(x) = n\right\} \\
&= (1 - D[1])^n \mathbb{P}\left\{\Phi_1 \mid |\mathcal{Y}_1| = n\right\} = (1 - D[1])^n A_{1;n},
\end{aligned}$$

implying (9a).

By a similar reasoning for $i > 1$,

$$\begin{aligned}
&\mathbb{P}\left\{\Phi_1; \dots; \Phi_i; |\mathcal{Y}_i| = n \mid \xi(x) = n\right\} \\
&= \mathbb{P}\left\{|\mathcal{Y}_i| = n \mid \xi(x) = n\right\} \cdot \mathbb{P}\left\{\Phi_1; \dots; \Phi_i \mid |\mathcal{Y}_i| = n; \xi(x) = n\right\} \quad (17) \\
&= (1 - D[i])^n A_{i;n}
\end{aligned}$$

Let \mathcal{X}_i denote the set of individuals that survive in the lineage xx_i : $\mathcal{X}_i = \{j : X_{i,j} \neq 0\}$. We rewrite the left-hand side of (17) by conditioning on the set of individuals that survive in lineage xx_i but not in the lineages xx_1, \dots, xx_{i-1} .

$$\begin{aligned}
&\mathbb{P}\left\{\Phi_1; \dots; \Phi_i; |\mathcal{Y}_i| = n \mid \xi(x) = n\right\} \\
&= \sum_{\mathcal{S} \in 2^{[n]}} \mathbb{P}\left\{\Phi_i; \mathcal{X}_i \setminus \mathcal{Y}_{i-1} = \mathcal{S} \mid \xi(x) = n\right\} \\
&\quad \times \mathbb{P}\left\{\Phi_1; \dots; \Phi_{i-1}; \mathcal{Y}_{i-1} = [n] \setminus \mathcal{S} \mid \xi(x) = n\right\} \\
&= \sum_{\substack{s+t=n \\ s+t=n}} B_{i;t,s} A_{i-1;t} \mathbb{P}\left\{|\mathcal{Y}_{i-1}| = t \mid \xi(x) = n\right\} \\
&= \sum_{\substack{s+t=n \\ s+t=n}} B_{i;t,s} A_{i-1;t} \binom{t+s}{s} (D[i-1])^s (1 - D[i-1])^t.
\end{aligned}$$

Combining this latter equality with (17) leads to (9b). \square

Proof of Theorem 6. By (4), Lemma 8 with $\sigma = D_x$ shows the relationship between the generating functions for the distributions of $\xi(x)$ and $\Xi(x)$. The theorem considers the case when x is the root and Corollary 9 applies to $\gamma(n) = \mathbb{P}\{\xi(x) = n\}$. \square

Proof of Theorem 7. By Theorem 5, the algorithm correctly computes the conditional survival likelihoods. Let T be the phylogeny with root ρ and n nodes. In order to prove the running time result, consider first the loop of Line 1. Lines 2–5 take $O(1)$ time for each $x \in \mathcal{V}(T)$. Line 8 is executed $(M_x + 1)^2$ times for every non-root x . If x is an inner node with children x_1, x_2, \dots, x_c , then

$M_x = \sum_{j=1}^c M_{x_j}$. Consequently,

$$\sum_{j=1}^c (M_{x_j} + 1)^2 \leq (M_x + 1)^2 + (c - 1). \quad (18)$$

Now, consider the *tree levels* $\mathcal{V}_0, \mathcal{V}_h$, where $\mathcal{V}_0 = \{\rho\}$, and for all $i = 1, \dots, h$, \mathcal{V}_i consists of all the children of nodes in \mathcal{V}_{i-1} . In other words, \mathcal{V}_i is the set of nodes that are reached through i edges from the root. By (18),

$$\sum_{y \in \mathcal{V}_i} (M_y + 1)^2 \leq |\mathcal{V}_i| - |\mathcal{V}_{i-1}| + \sum_{x \in \mathcal{V}_{i-1}} (M_x + 1)^2.$$

for all $i > 0$. Therefore,

$$\sum_{y \in \mathcal{V}_i} (M_y + 1)^2 \leq (|\mathcal{V}_i| - 1) + (M_\rho + 1)^2 \quad (19)$$

So,

$$\begin{aligned} \sum_{x \in \mathcal{V}(T) \setminus \{\rho\}} (M_x + 1)^2 &= \sum_{i=1}^h \sum_{x \in \mathcal{V}_i} (M_x + 1)^2 \\ &\leq n - 1 - h + h(M_\rho + 1)^2 \\ &= O(n + hM^2). \end{aligned}$$

Therefore, executing Line 8 through all iterations takes $O(M^2h + n)$ time. In order to bound the loop's running time in Line 9, consider the $B_{i;t,s}$ and $A_{i;n}$ values that are needed for a given node x with children x_1, \dots, x_c . By (8a), computing all $B_{i;0,s}$ values takes $O((M_{x_i} + 1)^2)$ time. Every $B_{i;t,s}$ with $t > 0$ and $A_{i;n}$ is calculated in $O(1)$ time. Using the bound $M[i] \leq M_x$, iteration i of the loop in Line 15 takes $O((M_x + 1)(M_{x_i} + 1))$ time. By summing for $i = 1, \dots, c$, we get that for node x with c_x children, the loop of Line 9 takes $O((M_x + 1)(M_x + c_x))$ time (since $\sum_i (M_{x_i} + 1) = M_x + c$). Now, $(M_x + 1)(M_x + c_x) = (M_x + 1)(M_x + 1 + c - 1)$, and

$$\begin{aligned} \sum_{x \in \mathcal{V}(T)} (M_x + 1)(c_x - 1) &= \sum_{i=0}^h \sum_{x \in \mathcal{V}_i} (M_x + 1)(c_x - 1) \\ &\leq (\max_x c_x - 1) \sum_{i=0}^h \sum_{x \in \mathcal{V}_i} (M_x + 1) \\ &\leq (c^* - 1)(M_\rho(h + 1) + n). \end{aligned}$$

By our previous discussions,

$$\begin{aligned} \sum_{x \in \mathcal{V}(T)} (M_x + 1)(M_x + c_x) &\leq n - 1 - h + h(M_\rho + 1)^2 + (c^* - 1)(M_\rho(h + 1) + n) \\ &= O(M^2h + c^*(Mh + n)). \end{aligned}$$

So, the loop of Line 9 takes $O(M^2h + Mhc^*)$ time, which leads to the Theorem's claim when combined with the bound on the loop's running time in Line 1. \square

Part III

Supplemental Methods

III.1 Universal conserved proteins

Table 7: Universal conserved proteins. The header lines give the COG functional category (J=translation, K=transcription, O=posttranslational modification, L=replication, U=trafficking, R=poorly characterized) in brackets, and the arCOG definition. For each family, the table lists the 28 genes that were used in the alignments by Genbank's gi identifier.

arCOG identifier	Description
arCOG00035	[R] Predicted ATPase of PP-loop superfamily
14600696 11499256 55377496 15759644 15678460 91772509 15668750 45357906 20094082 20090235 21228461	
84489373 88602856 41615083 48478039 18313507 14521349 18977200 70605951 15897585 57642215 16081635	
13541832 118194137 126353734 124027938 126179580 119720014	
arCOG00078	[J] Fibrillarin-like rRNA methylase
14601907 11499669 55377702 15790248 15679226 91773524 15668878 45358160 20093998 20089239 21227696	
84490205 88603510 41614921 48478284 18313923 14520277 18976431 70607104 15897821 57640118 16081994	
13541925 118194273 126354905 124026934 126178492 119719294	
arCOG00358	[R] Predicted GTPase
118431703 11499729 55378196 15790199 15679616 91773505 15669516 45359006 20094635 20093139 21227149	
84489469 88601924 41614908 48477308 18312780 14520715 18978135 70606742 15899138 57641221 16082451	
13540832 118194067 126354044 124027285 126179643 119719588	
arCOG00402	[J] Prolyl-tRNA synthetase
14601997 11499201 55377088 15759652 15678639 91772258 15669423 45358259 20094665 20092682 21226809	
84489803 88602429 41615004 48478436 18313148 14521178 18977665 70607289 15897489 57640485 16081983	
13541912 118194087 126353254 124027657 126179985 119719726	
arCOG00403	[J] Seryl-tRNA synthetase
118431673 11499617 55379420 15790918 15679133 91774249 15669265 45358442 20094896 20092841 21226967	
84490065 88603748 41615098 48478304 18313866 14521535 18977576 70607320 15897519 57641075 16081587	
13541818 118193827 126354149 124028106 126178761 119719325	
arCOG00410	[J] Phenylalanyl-tRNA synthetase alpha subunit
118431822 11499537 55379555 15791263 15678768 91773396 15668664 45359059 20093711 20089069 21227572	
84489226 88602293 41615216 48477197 18312639 14520329 18977361 70607254 15897061 57640856 16081727	
13541868 118193806 126354448 124027761 126178593 119719858	
arCOG00470	[L] ATPase involved in DNA replication HolB; large subunit
118431492 11498795 55378501 15790579 15678268 91774163 15669074 45357885 20093445 20090661 21226355	
84489226 88602207 41615216 48477945 18312141 14520329 18976464 70606692 15897671 57642154 16082282	
13541365 118195354 126354114 124027775 126178781 119719184	
arCOG00487	[J] Arginyl-tRNA synthetase
14601607 11498499 55379731 16120221 15679444 91773687 15668412 45358589 20094195 20088942 21227450	
84490352 88604070 41615001 48477675 18314006 14520895 18977752 70607149 15897753 57641170 16081423	
13542151 118194065 126354007 124028037 126180153 119719478	
arCOG00501	[R] Metal-dependent hydrolase of the beta-lactamase superfamily
118431611 11498544 55379160 15791057 15679819 91774034 15669696 45358469 20094469 20091849 21226408	
84489223 88602980 41614860 48478345 18313915 14521306 18977717 70607049 15897910 57641049 16082168	
13542061 118195558 126353577 124027542 126179500 119719606	
arCOG00675	[K] DNA-directed RNA polymerase; subunit E'
14600570 11498717 55379439 15790905 15678292 91774198 15668573 45358003 20094887 20092492 21226698	
84489433 88604096 41615160 48477938 18314155 14521883 18976628 70606624 15897347 57641634 16082590	
13541317 118194677 126353237 124027330 126180182 119719332	
arCOG00779	[J] Ribosomal protein L15
14600647 11499487 55378364 15790654 15678056 91772105 15668654 45358984 20093467 20089964 21228248	
84489685 88603477 41615106 48477334 18314064 14520535 33359574 70606390 15897602 57641454 16082253	
13541178 118194044 126354347 124028177 126178536 119719152	
arCOG00780	[J] Ribosomal protein L18E
14601599 11498727 55377000 15790217 15678068 91774297 15668365 45358886 20094911 20089484 21227857	
84489667 88604131 41615272 48477394 18312092 14520747 18978018 70605934 15897037 57641437 16081556	
13541965 118195195 126354325 124027416 126179747 119719197	
arCOG00781	[J] Ribosomal protein L32E
14600653 11499492 55378369 15790649 15678051 91772100 15668649 45358979 20094661 20089959 21228243	
84489690 88603482 41615313 48477729 18313096 14520540 18978179 70606395 15897607 57641459 16082256	
13541173 118194056 126353208 124028172 126178531 119719157	

Table 7 (continued). Universal conserved proteins.

arCOG identifier	Description
arCOG00785	[J] Ribosomal protein L29 118431014 11499502 55378377 15790639 15678040 91772090 15668639 45358968 20094279 20089949 21228233 84489701 88603492 41615057 4847719 18312872 33356679 18978190 70606405 15897616 57641470 16082663 13541162 118194579 126354084 124028162 126178521 119719403
arCOG00808	[J] Valyl-tRNA synthetase 118431600 11499806 55378488 15791298 15678792 91774107 15669196 45358147 20094860 20091870 21226416 84489310 88603484 41615045 48477604 18313245 14521851 18976662 70607182 15897784 57641209 16081220 13540863 118194774 126353043 124027752 126179518 119719242
arCOG00980	[O] FKBP-type peptidyl-prolyl cis-trans isomerase 2 14600774 11499571 55378030 15790337 15678273 15669012 45358135 20094667 20091637 21228758 84489146 88602455 41614799 48477400 18312191 14520961 18977773 70606452 15897659 57640376 16082046 13541409 118195252 126354604 124028403 126178885 119719223
arCOG00985	[J] Predicted RNA-binding protein (contains PUA domain) 118431091 11499000 55378796 15791374 15678677 91772255 15669623 45357618 20093659 20092679 21226806 8448968 88604286 41615000 48477515 18312763 14521517 18977487 70606790 15899938 57642210 16082395 13542272 118193986 126353971 124028415 126178182 119719227
arCOG00987	[J] Pseudouridine synthase 118431292 11497854 55379888 15790662 15678063 91772110 15668320 45359249 20093573 20089974 21228256 84489672 88603514 41615239 48477339 18312233 14520738 18978157 70606602 15897325 57641444 16082248 13541183 118194039 126353055 124028184 126178488 119719378
arCOG00991	[O] Prefoldin; molecular chaperone implicated in de novo protein folding; alpha subunit 14601786 11498315 55377126 15789666 15678204 91774129 15668612 45358173 20094392 20089020 21227510 84489330 88603521 41615094 48477526 18312177 14521321 18977418 70606464 15897215 57640695 16082455 13542328 118195434 126353730 124027360 126179629 119719616
arCOG01001	[J] Methionine aminopeptidase 118431332 11499425 55379957 15790763 15679298 91773582 15669519 45359007 20094051 20089110 21227601 84490276 88603584 41615186 48478163 18313107 14521622 18976913 70607256 15897059 57641118 16082408 13540934 118194936 126353044 124027759 126180032 119719590
arCOG01016	[K] DNA-directed RNA polymerase; subunit F (rpoF) 118431104 11499125 55377172 15790242 15679324 91772909 15668209 45357655 20095095 20090303 21228639 84490321 88604255 41615197 48477460 18313999 14521311 18977408 70606440 15897651 57640836 16082290 13541144 118194183 126353243 124028498 126179903 119719269
arCOG01183	[L] Iron containing AP-endonuclease 14601201 11498712 55379151 15790900 15679424 91773177 15669317 45357978 20094894 20092505 21226704 84488844 88604010 41615276 48477446 18313340 14521970 18976544 70606641 15897363 57642061 16081457 13542107 118195342 126354464 124027325 126180187 119719369
arCOG01344	[J] Ribosomal protein S19E (S16A) 14601178 11499651 55379741 15790874 15679611 91772194 15668873 45357719 20095056 20092910 21226904 84489403 88604243 41614983 48477275 18313782 14521029 18977871 70607213 15897290 57641211 16081227 13540838 118195458 126353307 124028439 126179916 119719387
arCOG01358	[J] 2-methylthioadenine synthetase 118431496 11498846 55379963 15790759 15678846 91773329 15669058 45357975 20094517 20090019 21228284 84489190 88603600 41614804 48477501 18313702 14521928 18978284 70606747 15899133 57641999 16081399 13542169 118194655 126354020 124027960 126178768 119719528
arCOG01560	[J] Translation initiation factor 2 (IF-2; GTPase) 14602026 11498374 55379763 15790864 15678287 91772291 15668436 45357847 20095031 20090384 21228565 84489438 88603730 41615279 48478356 18312680 14521344 18977509 70606498 15897175 57641240 16082134 13541284 118194017 126354606 124027334 126178411 119719414
arCOG01563	[J] Translation initiation factor 2; gamma subunit (eIF-2gamma; GTPase) 118431851 11498200 55379437 15790907 15678289 91774196 15669447 45358771 20094883 20092490 21226696 84489436 88604094 41615062 48477664 18311686 14520683 18978089 70606622 15897345 57641881 16081456 13542105 118195625 126353733 124027332 126180180 119719477
arCOG01722	[J] Ribosomal protein S13 118431572 11499860 55376996 15790213 15678064 91772118 15668361 45358882 20094907 20089977 21228257 84489671 88604135 41615251 48478291 18313810 14520743 18978022 70605930 15897041 57641441 16082067 13541393 118195300 126354556 124027412 126179743 119719177
arCOG01741	[R] Predicted RNA-binding protein 118431599 11498443 55377792 15790498 15679637 91773363 15668346 45357648 20093478 20089539 21227917 84489875 88601741 41614904 48477261 18313964 14520894 18977751 70605922 15897048 57640899 16082120 13541294 118194730 126353378 124027768 126178887 119719470
arCOG01751	[J] Ribosomal protein HS6-type (S12/L30/L7a)

Table 7 (continued). Universal conserved proteins.

arCOG identifier	Description
14601646 11498370 55377035 15790233 15678283 91772287 15669389 45358204 20095034 20090380 21228569 84489442 88602873 41615108 48478352 18314009 14520882 18977739 70607262 15897054 57641246 16082138 13541280 118194013 126353391 124027428 126178671 119719408	
arCOG01920	[K] Transcription antiterminator NusG
118431766 11498148 55377551 15789631 15679672 91773883 15668548 45358997 20094259 20093062 21227112 84490055 88601955 41615322 48477517 18313824 14520220 18978362 70607204 15897281 57641354 16082397 13542274 118194061 126354076 124028448 126178361 119719322	
arCOG01923	[J] Protein implicated in ribosomal biogenesis; Nop56p homolog
118431777 11499670 55377701 15790247 15679225 91773525 15668875 45358159 20093997 20089238 21227695 84490204 88603511 41615131 48477470 18313924 14520278 18976432 70607105 15897820 57640119 16082600 13541184 118194272 126354880 124026933 126178491 119719293	
arCOG01946	[J] Ribosomal protein S6E (S10)
118431852 11498122 55378902 15791272 15678288 91772292 15669446 45358770 20094882 20090385 21228564 84489437 88603729 41614902 48477663 18312675 14521633 18976860 70606621 15897344 57641886 16082513 13542106 118195624 126353077 124027333 126178412 119719339	
arCOG01988	[J] Translation elongation factor EF-1beta
118431890 11498182 55377174 15790245 15679693 91773295 15668636 45358964 20093695 20090330 21228616 84489951 88603589 41615014 48478236 18312109 14520246 18978337 70606568 15897128 57640047 16081669 13541449 118194239 126354516 124028209 126180383 119719571	
arCOG03013	[L] Eukaryotic-type DNA primase; large subunit
14600877 11497948 55379451 15791070 15678614 91774115 15668882 45357572 20094830 20089008 21227498 84488914 88602764 41615183 48478054 18313202 33356669 33359449 70607279 15897479 57641725 16082072 13541388 118195617 126354210 124027970 126179408 119719535	
arCOG04067	[J] Ribosomal protein L2
14600538 11499506 55378381 15790635 15678036 91772086 15668351 45359109 20093850 20089945 21228229 84489705 88603496 41615150 4847715 18312189 14520555 18978194 70606409 15897620 57641474 16082267 13541158 118195658 1263545873 124028148 126178517 119719145	
arCOG04070	[J] Ribosomal protein L3
14600545 11499509 55378384 15790632 15678033 91772083 15668348 45359106 20093853 20089942 21228226 84489708 88603499 41615218 4847712 18313001 14520558 18978197 70606412 15897623 57641477 16082270 13541155 118194573 126354760 124028157 126178514 119719148	
arCOG04071	[J] Ribosomal protein L4
118430943 11499508 55378383 15790633 15678034 91772084 15668349 45359107 20093852 20089943 21228227 84489707 88603498 41614941 4847713 18313002 14520557 18978196 70606411 15897622 57641476 16082269 13541156 118194574 126354761 124028158 126178515 119719147	
arCOG04072	[J] Ribosomal protein L23
118430944 11499507 55378382 15790634 15678035 91772085 15668350 45359108 20093851 20089944 21228228 84489706 88603497 41614861 4847714 18313003 14520556 18978195 70606410 15897621 57641475 16082268 13541157 118194575 126354763 124028159 126178516 119719146	
arCOG04077	[K] DNA-directed RNA polymerase; subunit E"
118430956 11498716 55379440 15790904 15678293 91774199 15668572 45358004 20094888 20092493 21226699 84489432 88604097 41615025 48477939 18314156 33356820 18976627 70606625 15897348 57641633 16082113 13541316 118194676 126354332 124027329 126180183 119719331	
arCOG04086	[J] Ribosomal protein L30
118431008 11499488 55378365 15790653 15678055 91772104 15668653 45358983 20093468 20089963 21228247 84489686 88603478 41615101 48477733 18312459 14520536 18978175 70606391 15897603 57641455 16082601 13541177 118194045 126353964 124028176 126178535 119719153	
arCOG04087	[J] Ribosomal protein S5
14600650 11499489 55378366 15790652 15678054 91772103 15668652 45358982 20093469 20089962 21228246 84489687 88603479 41615177 48477732 18312458 14520537 18978176 70606392 15897604 57641456 16082254 13541176 118194046 126353434 124028175 126178534 119719154	
arCOG04088	[J] Ribosomal protein L18
14600651 11499490 55378367 15790651 15678053 91772102 15668651 45358981 20093470 20089961 21228245 84489688 88603480 41614871 48477731 18313098 14520538 18978177 70606393 15897605 57641457 16082602 13541175 118194047 126353210 124028174 126178533 119719155	
arCOG04089	[J] Ribosomal protein L19E
14600652 11499491 55378368 15790650 15678052 91772101 15668650 45358980 20093471 20089960 21228244 84489689 88603481 41615168 48477730 18313097 14520539 18978178 70606394 15897606 57641458 16082255 13541174 118194057 126353209 124028173 126178532 119719156	
arCOG04090	[J] Ribosomal protein L6P
14600654 11499493 55378370 15790648 15678050 91772099 15668648 45358978 20094660 20089958 21228242 84489691 88603483 41615035 48477728 18313300 14520541 18978180 70606396 15897608 57641460 16082257 13541172 118194588 126353401 124028171 126178530 119719521	
arCOG04091	[J] Ribosomal protein S8

Table 7 (continued). Universal conserved proteins.

arCOG identifier	Description
118431009 11499494 55378371 15790647 15678049 91772098 15668647 45358977 20094659 20089957 21228241 84489692 88603484 41615066 48477727 18313095 14520542 18978181 70606397 15897609 57641461 16082258 13541171 118194587 126353499 124028170 126178529 119719158	
arCOG04092	[J] Ribosomal protein L5
118431011 11499496 55378372 15790645 15678047 91772096 15668646 45358975 20094657 20089955 21228239 84489694 88603486 41614891 48477725 18314167 14520544 18978183 70606399 15897611 57641463 16082260 13541169 118194585 126353408 124028168 126178527 119719160	
arCOG04093	[J] Ribosomal protein S4E
118431012 11499497 55378373 15790644 15678046 91772095 15668645 45358974 20094656 20089954 21228238 84489695 88603487 41615262 48477724 18313980 14520545 18978184 70606400 15897612 57641464 16082261 13541168 118194584 126353407 124028167 126178526 119719161	
arCOG04094	[J] Ribosomal protein L24
14600658 11499498 55378374 15790643 15678045 91772094 15668644 45358973 20094655 20089953 21228237 84489696 88603488 41615050 48477723 18313979 14520546 18978185 70606401 15897613 57641465 16082261 13541167 118194583 126353406 124028166 126178525 119719162	
arCOG04095	[J] Ribosomal protein L14
14600659 11499499 55378375 15790642 15678044 91772093 15668643 45358972 20094654 20089952 21228236 84489697 88603489 41614889 48477722 18313850 14520547 18978186 70606402 15897614 57641466 16082262 13541166 118194582 126353952 124028165 126178524 119719400	
arCOG04097	[J] Ribosomal protein S3
14600662 11499503 55378378 15790638 15678039 91772089 15668638 45358967 20094278 20089948 21228232 84489702 88603493 41615265 48477718 18312876 14520552 18978191 70606406 15897617 57641471 16082264 13541161 118194578 126354128 124028161 126178520 119719404	
arCOG04098	[J] Ribosomal protein L22
14600663 11499504 55378379 15790637 15678038 91772088 15668637 45358966 20094277 20089947 21228231 84489703 88603494 41614997 48477717 18312875 14520553 18978192 70606407 15897618 57641472 16082265 13541160 118194577 126353301 124028160 126178519 119719405	
arCOG04099	[J] Ribosomal protein S19
118431015 11499505 55378380 15790636 15678037 91772087 15668352 45359110 20094426 20089946 21228230 84489704 88603495 41615264 48477716 18312838 14520554 18978193 70606408 15897619 57641473 16082266 13541159 118194576 126354544 124028149 126178518 119719143	
arCOG04107	[J] Translation initiation factor 2; alpha subunit (eIF-2alpha)
118431043 11498138 55380037 15789766 15679308 91773369 15668289 45359270 20093857 20089531 21227909 84490305 88602418 41615191 48477477 18313775 14521054 18977512 70607040 15897919 57641035 16082212 13541243 118195638 126353921 124028479 126179994 119719552	
arCOG04108	[J] Ribosomal protein S27E
14600707 11498932 55380038 15789767 15679309 91773368 15668425 45359271 20093856 20089532 21227910 84490306 88602419 41615013 48477478 18313776 33356824 18976590 70607041 15897918 57641034 16082213 13541244 118195393 126353008 124028480 126179993 119719553	
arCOG04109	[J] Ribosomal protein L44E
118431044 11498931 55380039 15789768 15679310 91773367 15668424 45359272 20093855 20089533 21227911 84490307 88602420 41614997 48477479 18313777 14521914 33359458 70607042 15897917 57641033 16082214 13541245 118195685 126353919 124028481 126179992 119719554	
arCOG04111	[K] DNA-directed RNA polymerase; subunit L
118431051 11497823 55378181 15790001 15679317 91773407 15668563 45357824 20093498 20089606 21227982 84490313 88603581 41614976 48477485 18313007 14520268 18976422 70606015 15897236 57641102 16082389 13542266 118195124 126353726 124028487 126180108 119719281	
arCOG04112	[J] Diphthamide synthase subunit DPH2
118431053 11499391 55377297 15789998 15679319 91773404 15668660 45357751 20094953 20089603 21227979 84490315 88603851 41615020 48477502 18313748 14521333 18977435 70606017 15897238 57641863 16081398 13542170 118195704 126353270 124028491 126179789 119719791	
arCOG04113	[J] Ribosomal protein L16
14600720 11498937 55379031 15789424 15679130 91773391 15668723 45358852 20093691 20089081 21227578 84489148 88601407 41615235 48477787 18314160 14521607 18977651 70606018 15897239 57641481 16082585 13541370 118194845 126354059 124028492 126178798 119719460	
arCOG04116	[R] ATPase (PilT family)
14600730 11499533 55378805 15791349 15678274 91773737 15669728 45357844 20094972 20089787 21228122 84489459 88604155 41614855 48478233 18313913 14521315 18977413 70607293 15897493 57640888 16081452 13541271 118194908 126354780 124028133 126179950 119719450	
arCOG04129	[J] Ribosomal protein L21E
118431103 11499124 55377173 15790243 15679323 91772910 15668210 45357656 20095096 20090305 21228638 84490320 88604254 41615184 48477461 18313998 14521310 18977407 70606441 15897652 57640837 16082289 13541143 118194182 126353275 124028499 126179904 119719272	
arCOG04150	[R] Predicted RNA-binding protein (contains KH domains)

Table 7 (continued). Universal conserved proteins.

arCOG identifier	Description
118431180 11499393 55379140 15791048 15679024 91772948 15668619 45358168 20093951 20089777 21228114 84490126 88603706 41614827 48477323 18314094 14520809 18977952 70606751 15899129 57640735 16081522 13542007 118194496 126354047 124028254 126180438 119719426	
arCOG04154	[J] Ribosomal protein S8E
14600970 11499735 55378288 15790616 15678235 91774033 15668854 45359221 20094981 20089801 21228141 84490068 88603726 41615253 48478056 18313995 14521337 18977441 70606557 15897114 57641126 16082073 13541386 118195114 126353630 124027344 126178419 119719275	
arCOG04161	[J] Diphthamide biosynthesis methyltransferase
14601080 11497993 55378191 15790203 15679862 91772970 15669460 45358151 20094184 20090231 21228457 84489816 88602081 41615208 48478305 18312428 14521523 18976967 70607092 15897832 57640041 16081934 13541813 118193853 126353233 124027028 126179726 119719599	
arCOG04169	[U] Preprotein translocase subunit SecY
118431289 11499486 55378363 15790655 15678057 91772106 15668655 45358985 20093466 20089965 21228249 84489684 88603476 41614963 48477735 18314063 14520534 18978173 70606389 15897601 57641453 16082252 13541179 118194043 126354412 124028178 126178537 119719151	
arCOG04179	[R] DNA-binding protein
118431326 11499650 55379742 15790873 15679610 91772193 15668872 45357720 20095055 20092909 21226905 84489404 88604244 41614945 48477276 18313783 14521133 18977459 70607212 15897289 57641213 16081229 13540839 118195457 126353005 124028440 126179915 119719388	
arCOG04183	[J] Ribosomal protein S27AE
118431336 11498713 55379150 15790901 15678296 91774202 15668569 45358007 20094891 20092496 21226702 84489429 88604100 41614971 48477942 18313341 14520283 18976624 70606642 15897364 57641630 16082116 13541313 118194142 126354463 124027326 126180186 119719366	
arCOG04185	[J] Ribosomal protein S15P
14601204 11498407 55378209 15789952 15679422 91774257 15668206 45359142 20095029 20089827 21228168 84488841 88603752 41615271 48477316 18314103 14520283 18978428 70606619 15897342 57641186 16082150 13542039 118193830 126354459 124027322 126178380 119719429	
arCOG04186	[J] Ribosomal protein S3AE
118431343 11499901 55378212 15789949 15679588 91774248 15669170 45358232 20094897 20092228 21226286 84490085 88603749 41615164 48478596 18314093 14520285 18978426 70606435 15897646 57641189 16082179 13542078 118193826 126353494 124027319 126178377 119719563	
arCOG04187	[J] Predicted exosome subunit
118431347 11498102 55377979 15790325 15678712 91772265 15668772 45357813 20093822 20090629 21228723 84490037 88603529 41615307 48477698 18313180 14520824 18977942 70606427 15897638 57641571 16082606 13541136 118195242 126353010 124027454 126179621 119719496	
arCOG04208	[J] Ribosomal protein L37AE/L43A
14601410 11497677 55378437 15789534 15678708 91772269 15668773 45357812 20093817 20090626 21228727 84490041 88603530 41614833 48477464 18313176 14520490 18978380 70606423 15897634 57640550 16082287 13541140 118195238 126353662 124027458 126179620 119719406	
arCOG04209	[J] Ribosomal protein L15E
14601417 11499000 55378776 15789488 15678717 91772260 15669173 45357861 20093827 20090635 21228718 84490031 88603525 41614977 48477947 18312915 14521066 18977248 70606432 15897643 57641389 16082605 13541366 118194663 126353343 124027438 126179625 119719307	
arCOG04223	[J] Translation initiation factor 1 (eIF-1/SUI1)
14601525 11498519 55379110 15791324 15678041 91773983 15668640 45358969 20094280 20092852 21226960 84489700 88604144 41615056 48477220 18314004 14520550 18978189 70606655 15897376 57641469 16082603 13541163 118194918 126354287 124026956 126179940 119719420	
arCOG04228	[J] Peptidyl-tRNA hydrolase
14601543 11499678 55378525 16554461 15679691 91773256 15668222 45358172 20095033 20092085 21226214 84489948 88602102 41615039 48478175 18312120 14520837 18977928 70606566 15897126 57642235 16081275 13541018 118194234 126354596 124027644 126179346 119719594	
arCOG04239	[J] Ribosomal protein S4 or related protein
118431573 11499863 55376997 15790214 15678065 91772119 15668362 45358883 20094908 20089978 21228258 84489670 88604134 41615041 48478292 18313809 14520744 18978021 70605931 15897040 57641440 16082066 13541394 118195301 126353570 124027413 126179744 119719176	
arCOG04240	[J] Ribosomal protein S11
118431574 11499864 55376998 15790215 15678066 91772120 15668363 45358884 20094909 20089979 21228259 84489669 88604133 41614865 48478293 18313881 14520745 18978020 70605932 15897039 57641439 16082065 13541395 118194788 126353613 124027414 126179745 119719175	
arCOG04242	[J] Ribosomal protein L13
14601600 11498728 55377001 15790218 15678069 91774298 15668366 45358887 20094912 20089485 21227858 84489666 88604130 41615002 48477395 18312093 14520748 18978017 70605935 15897036 57641436 16081555 13541966 118195194 126353362 124027417 126179748 119719196	
arCOG04243	[J] Ribosomal protein S9
14601601 11498729 55377002 16554485 15678069 91774299 15668367 45358888 20094913 20089486 21227859 84489665 88604129 41615231 48477396 18312094 14520749 18978016 70605936 15897035 57641435 16081554 13541967 118195193 126353695 124027418 126179749 119719195	

Table 7 (continued). Universal conserved proteins.

arCOG identifier	Description
arCOG04244	[K] DNA-directed RNA polymerase; subunit N (RpoN/RPB10)
14601602 11498730 55377003 15790220 15678070 91774300 15668368 45358889 20094914 20089487 21227860 84489664 88604128 41615126 48477397 18312095 14520750 18978015 70605937 15897034 57641434 16081553 13541968 118195311 126353694 124027419 126179750 119719194	
arCOG04245	[J] Ribosomal protein S2
14601603 11498733 55377006 15790223 15678073 91774302 15669172 45358230 20095014 20089489 21227862 84489660 88604125 41615290 48477589 18312202 14520753 18978012 70605938 15897033 57641431 16082199 13541230 118195313 126353221 124027420 126179753 119719192	
arCOG04249	[J] tRNA nucleotidyltransferase (CCA-adding enzyme)
118431594 11499739 55378979 15789453 15678612 91773113 15669299 45358512 20094802 20092365 21226571 84488916 88604043 41614948 48477607 18313990 14520321 18976398 70607053 15897906 57641676 16082415 13540948 118195553 126354026 124026917 126178273 119719285	
arCOG04254	[J] Ribosomal protein S7
118431614 11499477 55379090 15791384 15679069 91773153 15669236 45358931 20094118 20090122 21228368 84490155 88604062 41615036 48477926 18312139 14520836 18977930 70606489 15897165 57641012 16081262 13540993 118194160 126353300 124028426 126180146 119719186	
arCOG04277	[J] Translation elongation factor P (EF-P)/translation initiation factor 5A (eIF-5A)
118431724 11498253 55378656 15790688 15678889 91773837 15669413 45358515 20094176 20092781 21227024 84490140 88602195 41615181 4847789 18314001 14520975 18977636 70607072 15897849 57640813 16082088 13541368 118194136 126354142 124028502 126178916 119719326	
arCOG04287	[J] Ribosomal protein L12E/L44/L45/RPP1/RPP2
14601888 11499087 55378201 15790194 15679676 91773887 15668685 45357821 20093444 20093066 21227116 84490051 88601951 41614996 48477512 18312824 14521981 18978366 70607200 15897277 57641350 16081487 13541251 118194947 126353544 124027292 126178365 119719787	
arCOG04288	[J] Ribosomal protein L10
118431764 11499086 55378200 15790195 15679675 91773886 15668686 45357822 20094262 20093065 21227115 84490052 88601952 41614888 48477511 18313827 14521982 18978365 70607201 15897278 57641351 16081488 13541252 118194955 126353495 124028329 126178364 119719140	
arCOG04289	[J] Ribosomal protein L1
118431765 11499085 55378199 15790196 15679674 91773885 15668687 45357823 20094261 20093064 21227114 84490053 88601953 41615328 48477510 18313826 14521983 18978364 70607202 15897279 57641352 16081489 13541253 118194956 126353304 124028450 126178363 119719141	
arCOG04302	[J] Glutamyl- or glutaminyl-tRNA synthetase
118431827 11497876 55377019 15790230 15678080 91774247 15669567 45358574 20094194 20089476 21227851 84489722 88604118 41615091 48477509 18313726 14520700 18978125 70607261 15897055 57641343 16081988 13541917 118195319 126353731 124027427 126179760 119719725	
arCOG04372	[J] Ribosomal protein L11
14601892 11498149 55378198 15790197 15679673 91773884 15668549 45358996 20094260 20093063 21227113 84490054 88601954 41614898 48477509 18313825 14520219 18978363 70607203 15897280 57641353 16081490 13541254 118194062 126354376 124028449 126178362 119719142	
arCOG04473	[J] Ribosomal protein L31E
118431324 11499648 55379680 15791237 15679607 91772190 15668220 45357625 20095053 20092906 21226908 84489406 88604247 41615154 48477278 18313818 14521719 33559475 70607210 15897287 57641255 16081231 13540841 118194653 126353007 124028442 126179912 119719390	

III.2 Likelihood correction for absent profiles

Suppose that profiles are restricted to the condition that $\{\Phi(x) > 0\}$ must hold for at least one terminal node x . The corresponding likelihoods

$$L_1 = \mathbb{P}\left\{\forall x \in \mathcal{L}(T): \xi(x) = \Phi(x) \mid \xi(x) > 0 \text{ for at least one leaf}\right\}$$

are obtained from the full likelihood by employing a correction that involves the probability of the condition (Felsenstein, 1992). Namely,

$$L_1 = \frac{L}{\mathbb{P}\{\xi(x) > 0 \text{ for at least one leaf}\}} = \frac{L}{1 - \mathbb{P}\{\xi(x) = 0 \text{ for all leaves } x\}}.$$

The probability that $\xi(x) = 0$ at all the leaves is the likelihood of the all-0 profile $\Phi_0 = (0, \dots, 0)$. By Theorem 5, $L_\rho[0] = \prod_{xy \in \mathcal{E}(T)} H_y(0)$ for the profile Φ_0 . Combined with (12), we have the correction formula $L_1 = \frac{L}{1-p_0}$ with

$$p_0 = \left(\prod_{xy \in \mathcal{E}(T)} H_y(0) \right) \cdot \exp\left(\Gamma(1 - D_\rho)\right), \quad (20)$$

for $\gamma(n) = \text{Poisson}(n; \Gamma)$.

III.3 Inferring family sizes at ancestors and counting lineage-specific events

Given a profile Φ , the posterior probabilities for gene family size at node x are computed by using the conditional survival likelihoods $L_x[n]$ and likelihoods of some relevant profiles on truncated phylogenies. In order to compute the gene content at node x , for example, consider the profile $\Phi_{x:m}$ for all m that applies to a phylogeny obtained by pruning the edges below x , that is, $\Phi_{x:m}(y) = \Phi(y)$ for $y \notin T_x$ and $\Phi_{x:m}(x) = m$. Let $L_{x:m}$ denote the likelihood of $\Phi_{x:m}$ on the pruned tree. Then

$$\mathbb{P}\left\{\xi(x) = m \mid \xi(\mathcal{L}(T)) = \Phi\right\} = \frac{L_{x:m} \sum_{n=0}^m \binom{m}{n} (D_x)^{m-n} (1-D_x)^n L_x[n]}{L}$$

gives the posterior probability that the family had m homologs at node x . In the analyses, we computed posterior probabilities for $m = 0, 1$, and used the complementary probability corresponding to multiple paralogs. The number of families present at node x , denoted by N_x , is inferred as a posterior mean value by summing posterior probabilities:

$$N_x = \sum_{i=1}^n \mathbb{P}\left\{\xi(x) > 0 \mid \xi(\mathcal{L}(T)) = \Phi_i\right\} + \frac{n \cdot p_0}{1-p_0} \mathbb{P}\left\{\xi(x) > 0 \mid \xi(\mathcal{L}(T)) = \Phi_0\right\}, \quad (21)$$

where $\Phi_i: i = 1, \dots, n$ are the profiles in the data set and p_0 is the likelihood of the all-0 profile Φ_0 from (20). Notice that the formula includes the absent all-0 profiles; there are $\frac{np_0}{1-p_0}$ such profiles by expectation.

Posterior probabilities of the general form $\mathbb{P}\left\{\xi(x) = n, \xi(y) = m \mid \xi(\mathcal{L}(T)) = \Phi\right\}$, characterizing lineage-specific family size changes on edge xy , can also be computed by using survival likelihoods on truncated phylogenies. In particular,

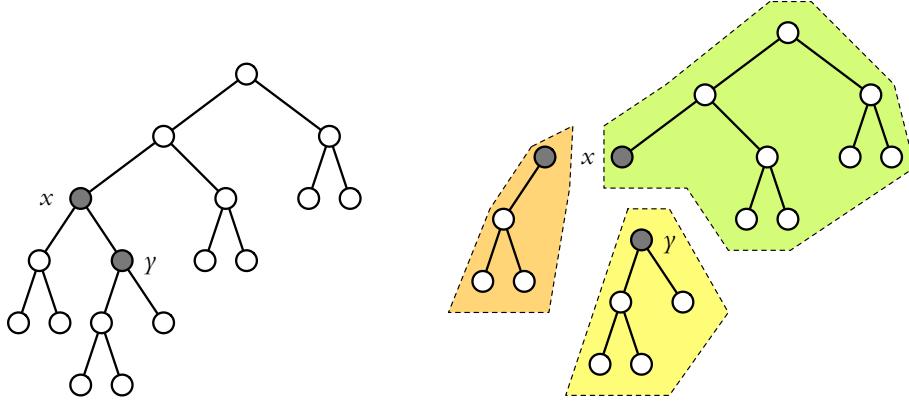


Figure 2: Decomposition of the phylogeny in order to compute posterior probability for lineage-specific events. Equation (22) shows the factors corresponding to the three parts.

we decompose the events as

$$\begin{aligned}
 \mathbb{P}\left\{\xi(x) = n, \xi(y) = m, \xi(\mathcal{L}(T)) = \Phi\right\} &= \mathbf{I} \times \mathbf{II} \times \mathbf{III} \\
 &= \mathbb{P}\left\{\xi(x) = n, \xi(\mathcal{L}(T) \setminus \mathcal{L}(T_x)) = \Phi(\mathcal{L}(T) \setminus \mathcal{L}(T_x))\right\} \\
 &\quad \times \mathbb{P}\left\{\xi(y) = m, \xi(\mathcal{L}(T_y)) = \Phi(\mathcal{L}(T_y)) \mid \xi(x) = n\right\} \\
 &\quad \times \mathbb{P}\left\{\xi(\mathcal{L}(T_x) \setminus \mathcal{L}(T_y)) = \Phi(\mathcal{L}(T_x) \setminus \mathcal{L}(T_y)) \mid \xi(x) = n\right\},
 \end{aligned} \tag{22}$$

where the second factor can be written as

$$\begin{aligned}
 \mathbb{P}\left\{\xi(y) = m, \xi(\mathcal{L}(T_y)) = \Phi(\mathcal{L}(T_y)) \mid \xi(x) = n\right\} \\
 = \mathbb{P}\left\{\xi(y) = m \mid \xi(x) = n\right\} \sum_{k=0}^m \binom{m}{k} (D_y)^{m-k} (1 - D_y)^k L_y[k].
 \end{aligned}$$

Figure 2 illustrates the decomposition of the phylogeny into three parts, corresponding to the three factors in (22).

For each edge xy , we computed the posterior probabilities for *gain*, *loss*, *expansion* and *contraction*:

$$\begin{aligned}
 \mathbb{P}\{\text{gain}(xy)\} &= \mathbb{P}\{\xi(x) = 0, \xi(y) > 0\} = \mathbb{P}\{\xi(x) = 0\} - \mathbb{P}\{\xi(x) = 0, \xi(y) = 0\} \\
 \mathbb{P}\{\text{loss}(xy)\} &= \mathbb{P}\{\xi(x) > 0, \xi(y) = 0\} = \mathbb{P}\{\xi(y) = 0\} - \mathbb{P}\{\xi(x) = 0, \xi(y) = 0\} \\
 \mathbb{P}\{\text{expansion}(xy)\} &= \mathbb{P}\{\xi(x) = 1, \xi(y) > 1\} = \mathbb{P}\{\xi(x) = 1\} \\
 &\quad - \mathbb{P}\{\xi(x) = 1, \xi(y) = 0\} - \mathbb{P}\{\xi(x) = 1, \xi(y) = 1\} \\
 \mathbb{P}\{\text{contraction}(xy)\} &= \mathbb{P}\{\xi(x) > 1, \xi(y) = 1\} = \mathbb{P}\{\xi(y) = 1\} \\
 &\quad - \mathbb{P}\{\xi(x) = 0, \xi(y) = 1\} - \mathbb{P}\{\xi(x) = 1, \xi(y) = 1\},
 \end{aligned}$$

where all probabilities are conditioned on the observation of the phylogenetic profile $\{\xi(\mathcal{L}(T)) = \Phi\}$. Expected numbers for gains, losses, expansions and contractions on each edge xy were computed by formulas analogous to (21).

III.4 Rate variation

In order to capture lineage- and family-specific rate variation, we employ models in which the total gain, loss, and duplication rates for a paralog family f on a branch e can be written as

$$\kappa \hat{t}_e = \hat{t}_e t_f \kappa_e \kappa_f, \quad \mu \hat{t}_e = \hat{t}_e t_f \mu_e \mu_f, \quad \lambda \hat{t}_e = \hat{t}_e t_f \lambda_e \lambda_f.$$

Lineage-specific rates $\hat{\kappa}_e$, $\hat{\mu}_e$ and $\hat{\lambda}_e$ are set numerically by fixing $\hat{t}_e = 1$ during optimization. Family-specific rate factors t_f , κ_f , μ_f and λ_f are either kept as constant 1, or are assumed to follow an approximation of a Gamma distribution with four discrete categories (Yang, 1994). In such a rate variation model, the likelihood of a profile Φ for family f is computed as

$$L = \sum_{c \in \mathcal{C}} \mathbb{P} \left\{ \xi(\mathcal{L}(T)) = \Phi \mid (t_f, \kappa_f, \mu_f, \lambda_f) = (t_c, \kappa_c, \mu_c, \lambda_c) \right\} \frac{1}{|\mathcal{C}|},$$

where \mathcal{C} denotes the set of family-specific rate categories, which is the product space for the rate factor distributions, and $t_c, \kappa_c, \mu_c, \lambda_c$ are computed as the means within the four quartiles of a Gamma distribution with mean 1. In order to account for absent profiles, likelihoods need to be corrected using the formula $L_1 = L/(1 - p_0)$ where the probability $p_0 = \sum_{c \in \mathcal{C}} \frac{p_0(c)}{|\mathcal{C}|}$ is computed using the category-specific probabilities $p_0(c)$ for all-0 profiles.

Bibliography

- Felsenstein, J. (1992). Phylogenies from restriction sites, a maximum likelihood approach. *Evolution*, **46**, 159–173.
- Guindon, S. and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, **52**(5), 696–704.
- Karlin, S. and McGregor, J. (1958). Linear growth, birth, and death processes. *Journal of Mathematics and Mechanics*, **7**(4), 643–662.
- Kendall, D. G. (1949). Stochastic processes and population growth. *Journal of the Royal Statistical Society Series B*, **11**(2), 230–282.
- Ross, S. M. (1996). *Stochastic Processes*. Wiley & Sons, New York, second edition.
- Takács, L. (1962). *Introduction to the Theory of Queues*. Oxford University Press, New York.
- Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution*, **39**, 306–314.